

ENGINEERING CYBERNETICS

ENGINEERING CYBERNETICS

H. S. TSIEN

Daniel and Florence Guggenheim Jet Propulsion Center
California Institute of Technology
Pasadena, California

McGRAW-HILL BOOK COMPANY, INC.

New York Toronto London

1954

ENGINEERING CYBERNETICS

Library of Congress Catalog Card Number 54-8098

THE MAPLE PRESS COMPANY, YORK, PA.

To

Tsiang Yin

PREFACE

The celebrated physicist and mathematician A. M. Ampère coined the word *cybernétique* to mean the science of civil government (Part II of "Essai sur la philosophie des sciences," 1845, Paris). Ampère's grandiose scheme of political sciences has not, and perhaps never will, come to fruition. In the meantime, conflict between governments with the use of force greatly accelerated the development of another branch of science, the science of control and guidance of mechanical and electrical systems. It is thus perhaps ironical that Ampère's word should be borrowed by N. Wiener to name this new science, so important to modern warfare. The "cybernetics" of Wiener ("Cybernetics, or Control and Communication in the Animal and the Machine," John Wiley & Sons, Inc., New York, 1948) is the science of organization of mechanical and electrical components for stability and purposeful actions. A distinguishing feature of this new science is the total absence of considerations of energy, heat, and efficiency, which are so important in other natural sciences. In fact, the primary concern of cybernetics is on the qualitative aspects of the interrelations among the various components of a system and the synthetic behavior of the complete mechanism.

The purpose of "Engineering Cybernetics" is then to study those parts of the broad science of cybernetics which have direct engineering applications in designing controlled or guided systems. It certainly includes such topics usually treated in books on servomechanisms. But a wider range of topics is only one difference between engineering cybernetics and servomechanisms engineering. A deeper—and thus more important—difference lies in the fact that engineering cybernetics is an engineering *science*, while servomechanisms engineering is an engineering *practice*. An engineering science aims to organize the design principles used in engineering practice into a discipline and thus to exhibit the similarities between different areas of engineering practice and to emphasize the power of fundamental concepts. In short, an engineering science is predominated by theoretical analysis and very often uses the tool of advanced mathematics. A glance at the contents of this book makes this quite evident. The detailed construction and design of the components of the system—the actual implementation of the theory—are almost never discussed. No gadget is mentioned.

What is the justification of this *separation* of the theory from the practice? With knowledge of the very existence of various engineering sciences and their recent rapid development, such justification seems hardly necessary. Moreover, a specific example could be cited: Fluid mechanics exists as an engineering science separate from the practice of aerodynamics engineers, hydraulic engineers, meteorologists, and many others who use the results of investigations in fluid mechanics in their daily work. In fact, without fluid mechanists, the understanding and the utilization of supersonic flows would certainly be greatly delayed, to say the least. Therefore, the justification of establishing engineering cybernetics as an engineering science lies in the possibility that looking at things in broad outline and in an organized way often leads to fruitful new avenues of approach to old problems and gives new, unexpected vistas. At the present stage of multifarious developments in control and guidance engineering, there is a very real advantage in trying to grasp the full potentialities of this new science by a comprehensive survey of the whole field.

Therefore a discussion on engineering cybernetics should cover reasonably well all aspects of the science expected to have engineering applications and, in particular, should not avoid a topic for the mere reason of mathematical difficulties. This is all the more true when one realizes that the mathematical difficulties of any subject are usually quite artificial. With a little reinterpretation, the matter could generally be brought down to the level of a research engineer. The mathematical level of this book is then that of a student who has had a course in elements of mathematical analysis. Knowledge of complex integration, variational calculus, and ordinary differential equations forms the prerequisite for the study. On the other hand, no rigorous and elegant mathematical argument is introduced if a heuristic discussion suffices. Hence to the practicing electronics specialist, the treatment here must appear to be excessively "long-hair," but to a mathematician interested in this field, the treatment here may well appear to be amateurish. If indeed these are the only criticisms, then, with all due respect to them, the author shall feel that he has not failed in what he aimed to do.

During the course of writing these chapters, the author had the benefit of many conversations with his colleagues at the California Institute of Technology, Dr. Frank E. Marble and Dr. Charles R. DePrima, which often led to sudden clarification of an obscure point. The task of preparing the manuscript was greatly lightened by the efficient help rendered by Sedat Serdengecti and Ruth L. Winkel. To all of them, the author wishes to extend his sincere thanks.

H. S. TSIEN

CONTENTS

ENGINEERING CYBERNETICS

CHAPTER 1

INTRODUCTION

Consider a system of one degree of freedom so that the physical state of the system can be specified by a single variable y. The behavior of the system is then described by taking y as a function of time t. To determine this behavior or $y(t)$, it is necessary to know the structure of the system and the properties of the individual elements of the system. This knowledge about the system, together with fundamental physical laws, when translated into mathematical language gives an equation for calculating the function $y(t)$. This equation could be an integral equation or an integrodifferential equation, but very often it is a differential equation. It is also an ordinary differential equation, because there is only one independent variable, the time t.

A differential equation is called linear, and the system described by the differential equation, a *linear system*, if each term of the equation contains at most only first powers of the dependent variable y or its time derivatives. The terms should not contain higher powers of y or cross products of y and its derivatives. Otherwise, the differential equation is called nonlinear, and the system described by the differential equation, a *nonlinear system*. Linear systems can be further subdivided into systems with constant coefficients and systems with variable coefficients. Constant-coefficient systems have constants independent of time t as coefficients of the terms in the differential equation describing the system. Variable-coefficient systems have coefficients that are functions of t.

This concern about the classification of the differential equation has its justification in that the character of the solution of the equation and hence the behavior of the system depend closely on the type of the differential equation which describes it. Even more than this, the type of differential equation specifies the kind of questions that can logically be asked about the system. In other words, the type of differential equation determines the proper approach to the solution of the engineering problem of the system. We shall see this presently.

1.1 Linear Systems of Constant Coefficients. Let us consider the simplest system—a first-order system. That is, the differential equation of the system is a first-order linear differential equation of constant coefficients. If the system is assumed to be free and is not subjected to

"forcing functions," then the differential equation can be written as

$$\frac{dy}{dt} + ky = 0 \tag{1.1}$$

k may be called the spring constant and is real. When there is no variation of y with respect to time, dy/dt vanishes, and Eq. (1.1) requires $y = 0$. Therefore the stationary, or equilibrium, state of the system corresponds to $y = 0$.

The solution of Eq. (1.1) is

$$y = y_0 e^{-kt} \tag{1.2}$$

where y_0 is the initial value of y, or

$$y(0) = y_0 \tag{1.3}$$

y_0 is thus the initial disturbance of the system from the equilibrium state. The behavior of the system for $t > 0$ is illustrated in Fig. 1.1 for both positive and negative k. It is seen that for $k > 0$ the magnitude of y decreases with time. Then, as the time increases indefinitely, $y \to 0$. Therefore, for $k > 0$, the disturbance of the system will eventually disappear. The system can then be said to be *stable*. When $k < 0$, the motion of the system increases with time, and eventually the disturbance will become very large no matter how small the initial displacement is: the system will never return to the equilibrium state once disturbed. Such systems are thus *unstable*.

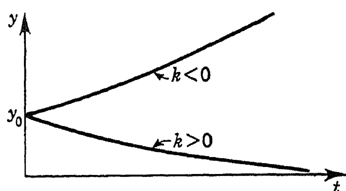


FIG. 1.1

For systems of higher order, the differential equation will have higher derivatives. The nth-order system has the differential equation

$$\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_0 y = 0 \tag{1.4}$$

For a physical system, the coefficients $a_{n-1}, \ldots, a_0$ are real. Then the solution of Eq. (1.4) can be written as

$$y = \sum_{i=1}^{n} y_0^{(i)} e^{\alpha_i t} \sin (\beta_i t + \varphi_i) \tag{1.5}$$

where α_i, β_i are real and are related to the coefficients $a_{n-1}, \ldots, a_0$, and the φ_i's are the phase angles. The motion of the system is thus stable only if all α_i's are negative. If one of them is positive, the disturbance will eventually diverge, and the system is thus unstable.

From the above examples it is seen that the crucial question to ask about the behavior of a linear system of constant coefficients is the question of stability. Needless to say, the usual aim of an engineering design is stability. The question of stability can be answered, however, once the coefficients of the differential equation are specified. In the case of the simple first-order system specified by Eq. (1.1), the only information that matters is the sign of the coefficient k.

1.2 Linear Systems of Variable Coefficients. If there is a variable parameter in the system under study, the stationary, or equilibrium, state of the system can be changed by changing this parameter. It is natural, then, to expect the coefficients of the linear differential equation describing the system to be also functions of this parameter. For instance, the aerodynamic forces acting on an aircraft are functions of the speed of the aircraft. If the speed of the aircraft is changing owing to acceleration or deceleration, the aerodynamic forces will change accordingly while the inertial properties of the aircraft remain practically the same. As a result, if we wish to calculate the disturbed motion of the aircraft from, say, horizontal flight, the fundamental differential equation will be an equation with variable coefficients.

Let us return to the simple example of a first-order system, described by Eq. (1.1). If the spring constant k is a function of the speed of the aircraft and if the aircraft has a constant acceleration a, then k is a function of the velocity $u = at$. Thus the differential equation can be written as

$$\frac{dy}{dt} + k(at)y = 0 \tag{1.6}$$

The solution of this equation is

$$\log \frac{y}{y_0} = -\frac{1}{a}\int_0^{at} k(\xi)\, d\xi \tag{1.7}$$

where y_0 is the initial disturbance. If k is always positive then $\log (y/y_0)$ is always negative, and as time increases $\log (y/y_0)$ will be increasingly negative. Therefore y is always less than y_0 and eventually will vanish. Thus the system is stable. If k is always negative, $\log (y/y_0)$ will be increasingly positive with time. Then y will eventually become very large even if the initial disturbance y_0 is very small. The system is thus unstable. These characteristics of the linear system with variable coefficients that remain positive or negative are very similar to those of systems with constant coefficients.

The interesting case is, however, when k has both positive and negative values. Let us assume $k(at)$ to be first positive, then negative, but finally positive again. If the first zero of k is denoted by $u_1 = at_1$

and the second zero denoted by $u_2 = at_2$, then according to our previous concepts, the system is unstable in the velocity range from u_1 to u_2 (Fig. 1.2). Let y_{min} be the minimum value of y and y_{max} be the maximum value of y. Then Eq. (1.7) gives

$$\log \frac{y_{min}}{y_0} = -\frac{1}{a}\int_0^{u_1} k(\xi)\, d\xi \tag{1.8}$$

and

$$\log \frac{y_{max}}{y_0} = -\frac{1}{a}\int_0^{u_2} k(\xi)\, d\xi \tag{1.9}$$

Of primary engineering interest is the question: How large is y_{max}? Is it so large that the system cannot function properly? We note that

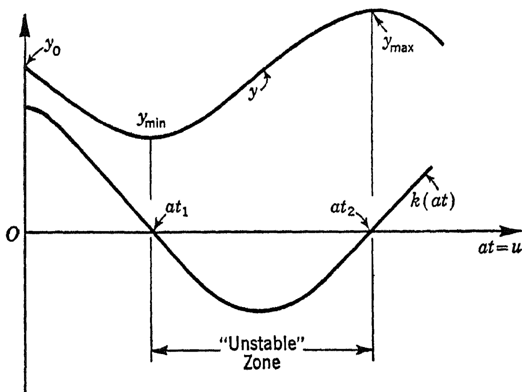


Fig. 1.2

to answer this question the knowledge of two things, in addition to the functional dependence of k upon u, is required. These are: How large is the acceleration a? What is the magnitude of the initial disturbance y_0? For any fixed a, y_{max} is proportional to y_0. But more important, for any fixed initial disturbance, the maximum value of the deviation y_{max} can be greatly reduced by increasing the acceleration a, as shown by Eq. (1.9). This means that by going through the "unstable" zone quickly, the undesirable effects can be minimized.

Therefore, for the more general linear systems with variable coefficients, the simple question of stability has no definite meaning. The more meaningful question is to ask specifically whether under a definite criterion the system will behave satisfactorily with specified disturbances and circumstances. In our simple example of a first-order system, the definite criterion of proper behavior is y_{max}; the specified disturbance is y_0, and the specified circumstance is the acceleration a. Thus by going from systems with constant coefficients to systems with variable coefficients, the character of the problem is already considerably modified.

1.3 Nonlinear Systems. If the spring constant k of the simple first-order system described by Eq. (1.1) is a function of the disturbance y itself, then the differential equation is

$$\frac{dy}{dt} + f(y) = 0 \tag{1.10}$$

where $f(y) = k(y)y$. It is seen that the differential equation is nonlinear. The system described by Eq. (1.10) is thus the simplest example of a nonlinear system. The solution $y(t)$ can be computed from the following relation obtained by integrating Eq. (1.10):

$$t = -\int_{y_0}^{y} \frac{d\eta}{f(\eta)} \tag{1.11}$$

where y_0 is again the initial disturbance.

On the other hand, repeated differentiation of Eq. (1.10) gives

$$\left.\begin{aligned} \frac{d^2y}{dt^2} + \frac{df}{dy}\frac{dy}{dt} &= 0 \\ \frac{d^3y}{dt^3} + \frac{d^2f}{dy^2}\left(\frac{dy}{dt}\right)^2 + \frac{df}{dy}\frac{d^2y}{dt^2} &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots \end{aligned}\right\} \tag{1.12}$$

Thus if y_1 is a zero of the function $f(y)$ and if $f(y)$ is regular at y_1 so that all the derivatives of $f(y)$ with respect to y are finite at y_1, then from Eqs. (1.10) and (1.12)

$$\frac{dy}{dt} = \frac{d^2y}{dt^2} = \frac{d^3y}{dt^3} = \cdots = 0 \qquad \text{at } y = y_1 \tag{1.13}$$

This means that y approaches y_1 asymptotically. In fact, if $y_0 > y_1$ and $f(y_0) > 0$, then y will become y_1 eventually. If $y_0 < y_1$, then $f(y_0) < 0$, and y will again become y_1 at $t \to \infty$. This pattern of behavior of y is repeated with other zeros of the $f(y)$ (Fig. 1.3).

If the initial disturbance y_0 coincides with one of the zeros of $f(y)$, this value of y will be maintained with increasing time. Thus the zeros of $f(y)$ are equilibrium positions. If $df/dy > 0$ at a zero such as y_1, small deviations from this equilibrium position will eventually disappear and the system will finally return to the initial state. Thus the system may be said to have stability for small disturbances at y_1. If, however, $df/dy < 0$ at a zero such as y_2, the slightest disturbance from this equilibrium position will cause the system to move to the next equilibrium positions y_1 or y_3. y_2 is thus an unstable equilibrium state.

We have seen that even for the very simple nonlinear system described by Eq. (1.10), the behavior of the system is very complicated. The

system may have both stability and instability. Thus for such systems, it is entirely senseless to ask a general question about stability; rather, each specific problem must be considered individually.

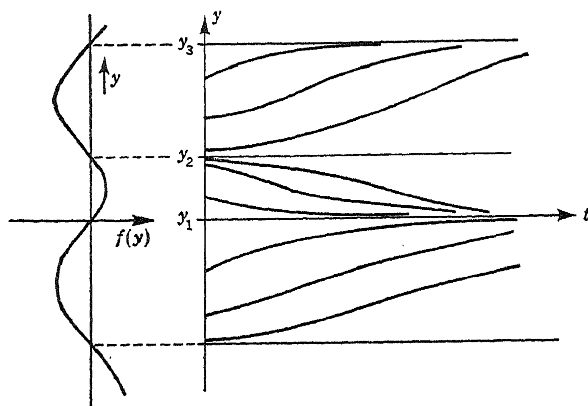


Fig. 1.3

1.4 Engineering Approximation. It is almost certain that any physical system, if analyzed in great detail, is always nonlinear. We speak of the system as linear only with the understanding that we mean the system can be approximated sufficiently accurately by a linear system. Furthermore, sufficient accuracy means that the deviation from linearity is so small as to be unimportant for the specific problem considered. Thus whether a system is linear or not can be determined only under clearly defined circumstances. There is no general absolute criterion.

The same can be said for the classification of linear systems into systems of constant coefficients and systems with variable coefficients. Take our simple examples described by Eqs. (1.1) and (1.6). If the acceleration a is very small, *i.e.*, flight at almost constant speed, Eq. (1.8) shows that $y_{\min}$ will be very much smaller than the initial disturbance y_0, and this value of $y_{\min}$ will occur at a very large value of t. The behavior of the system within a finite time interval is thus very similar to a system described by Eq. (1.1) with positive k. Therefore the system of variable coefficients can be sufficiently accurately approximated by a system with constant coefficients under certain circumstances.

Needless to say, linear systems with constant coefficients are the easiest to study. It is fortunate that a very large number of engineering systems falls into this classification when the "engineering approximation" is made. This is the reason why this particular field of stability and control theory is the most developed one. In fact, the present theory of servomechanism deals almost exclusively with such systems. We shall therefore begin with linear systems of constant coefficients.

CHAPTER 2

METHOD OF LAPLACE TRANSFORM

For linear differential equations with constant coefficients and with time t as the independent variable, the method using Laplace transforms is particularly useful in finding the solution. Of course the problem can be solved by a number of other methods; but the Laplace-transform method appeals especially to the engineering scientist in that it reduces all problems to a uniform basis. The procedure of solution is then standardized, and a general approach is possible. The theory and practice of the Laplace transform are discussed in many texts.[1] It is not the purpose of the present chapter to do this. The purpose here is rather to give for easy reference a summary of results which are useful to our discussion in the subsequent chapters. For details and proofs, the reader should consult the texts cited.

2.1 Laplace Transform and Inversion Formula. If $y(t)$ is a function of the time variable t defined for $t > 0$, the Laplace transform $Y(s)$ of $y(t)$ is defined as[2]

$$Y(s) = \int_0^\infty e^{-st} y(t)\, dt \tag{2.1}$$

where s is a complex variable having a positive real part, $\Re s > 0$. For other values of s, the function $Y(s)$ is defined by the analytic continuation. The dimension of $Y(s)$ is the dimension of y multiplied by time.

When $Y(s)$ is known, the original function for which $Y(s)$ is the Laplace transform can be obtained in all cases by the inversion formula:

$$y(t) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{st} Y(s)\, ds \tag{2.2}$$

[1] See for instance H. S. Carslaw and J. C. Jaeger, "Operational Methods in Applied Mathematics," Oxford University Press, New York, 1941; or R. V. Churchill, "Modern Operational Methods in Engineering," McGraw-Hill Book Company, Inc., New York, 1944. For a more complete theory, one should consult G. Doetsch, "Theorie und Anwendung der Laplace-Transformation," Verlag Julius Springer, Berlin, 1937; or D. V. Widder, "The Laplace Transform," Princeton University Press, Princeton, N. J., 1946.

[2] Throughout this book, capital letters will be used to denote the Laplace transform of quantities denoted by lower-case letters.

where γ is a constant greater than the real part of all the singularities of $Y(s)$. The actual evaluation of $y(t)$ can be done by properly deforming the path of integration according to the character of $Y(s)$.

2.2 Application to Linear Equations with Constant Coefficients. Since the Laplace transform is defined as an operation on a function defined for $t > 0$, the method is particularly adapted to initial-value problems: Given the initial state of the system and the forcing function for $t > 0$, find the "motion" of the system for $t > 0$. Let us consider an nth-order system, with coefficients $a_n, a_{n-1}, \ldots, a_0$ for the derivatives, and a nonhomogeneous term, or forcing function, $x(t)$. Then the differential equation is

$$a_n \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_0 y = x(t) \tag{2.3}$$

The initial conditions are usually specified as

$$\left.\begin{aligned} \left(\frac{d^{n-1}y}{dt^{n-1}}\right)_{t=0} &= y_0^{(n-1)} \\ \cdots\cdots&\cdots\cdots \\ (y)_{t=0} &= y_0 \end{aligned}\right\} \tag{2.4}$$

The differential equation (2.3) together with the condition (2.4) determines uniquely the behavior of the system for $t \geq 0$.

To solve the problem by the Laplace transformation, we multiply both sides of Eq. (2.3) by e^{-st} and integrate from $t = 0$ to $t = \infty$. Then

$$\int_0^\infty e^{-st} y(t)\, dt = Y(s) \tag{2.1a}$$

And by partial integration,

$$\left.\begin{aligned} \int_0^\infty e^{-st} \frac{dy}{dt}\, dt &= -y_0 + s \int_0^\infty e^{-st} y(t)\, dt = -y_0 + sY(s) \\ \int_0^\infty e^{-st} \frac{d^2y}{dt^2}\, dt &= -y_0^{(1)} - sy_0 + s^2 Y(s) \\ \cdots\cdots&\cdots\cdots\cdots\cdots \\ \text{and} \qquad & \\ \int_0^\infty e^{-st} \frac{d^ny}{dt^n}\, dt &= -y_0^{(n-1)} - sy_0^{(n-2)} - \cdots - s^{n-1}y_0 + s^n Y(s) \end{aligned}\right\} \tag{2.5}$$

Therefore, if the Laplace transform of the forcing function $x(t)$ is written as $X(s)$, *i.e.*,

$$X(s) = \int_0^\infty e^{-st} x(t)\, dt \tag{2.6}$$

then Eq. (2.3) together with the initial conditions (2.4) can be written as

$$(a_n s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0)Y(s) = a_n y_0 s^{n-1} + (a_n y_0^{(1)} + a_{n-1}y_0)s^{n-2} + (a_n y_0^{(2)} + a_{n-1}y_0^{(1)} + a_{n-2}y_0)s^{n-3} + \cdots + (a_n y_0^{(n-1)} + a_{n-1}y_0^{(n-2)} + \cdots + a_0 y_0) + X(s) \quad (2.7)$$

Hence if we define the polynomials $D(s)$ and $N_0(s)$ as

$$D(s) = a_n s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 \quad (2.8)$$

and

$$N_0(s) = a_n y_0 s^{n-1} + (a_n y_0^{(1)} + a_{n-1}y_0)s^{n-2} + \cdots + (a_n y_0^{(n-1)} + a_{n-1}y_0^{(n-2)} + \cdots + a_0 y_0) \quad (2.9)$$

then the solution of Eq. (2.7) is

$$Y(s) = \frac{N_0(s)}{D(s)} + \frac{X(s)}{D(s)} \quad (2.10)$$

We note that the first term of the solution given by Eq. (2.10) depends on the initial conditions through Eq. (2.9). $N_0(s)$ is at most of order $n - 1$ and is thus a lower order than $D(s)$. $N_0(s)$ will vanish if all the initial values specified by Eq. (2.4) vanish. In that case, $Y(s)$ is given by the second term alone. The second term depends upon the forcing function. Therefore the first term, $N_0(s)/D(s)$, can be called the complementary function, and the second term $X(s)/D(s)$, the particular integral. The actual solution $y(t)$ can be obtained from $Y(s)$ of Eq. (2.10) by applying the inversion formula of Eq. (2.2).

2.3 "Dictionary" of Laplace Transforms. The forcing function $x(t)$ is often of a character such that $X(s)$ is the ratio of two polynomials in s. Then the complete solution $Y(s)$ given by Eq. (2.10) is also the ratio of two polynomials of s. Therefore the expressions for $Y(s)$ can be broken down into a number of simple fractions. Each of the fractions can be inverted by the inversion formula, or, better yet, the original

TABLE 2.1
DICTIONARY OF LAPLACE TRANSFORMS

$Y(s)$	$y(t)$
$1/s$	1
$1/s^n$	$t^{n-1}/\Gamma(n)$
$1/s - a$	e^{at}
$a/(s^2 + a^2)$	$\sin at$
$s/(s^2 + a^2)$	$\cos at$
$a/(s^2 - a^2)$	$\sinh at$
$s/(s^2 - a^2)$	$\cosh at$
$s/(s^2 + a^2)^2$	$\frac{t}{2a}\sin at$
$1/(s^2 + a^2)^2$	$\frac{1}{2a^3}(\sin at - at\cos at)$

functions of t can be found by the use of a "dictionary," a list of simple functions of t and their Laplace transforms. A much abridged dictionary is given in Table 2.1.

2.4 Sinusoidal Forcing Function. The ratio of polynomials $N_0(s)/D(s)$ can be broken down into partial fractions. If the roots of the polynomial $D(s)$ are all different, say $s_1, s_2, \ldots, s_n$, then

$$\frac{N_0(s)}{D(s)} = \sum_{r=1}^{n} \frac{N_0(s_r)}{D'(s_r)} \frac{1}{(s - s_r)}$$

where $D'(s)$ is the derivative of $D(s)$ with respect to s. By "interpreting" the sum term by term according to our dictionary, the part $y_c(t)$ of the solution due to the initial conditions, or the complementary function, is

$$y_c(t) = \sum_{r=1}^{n} \frac{N_0(s_r)}{D'(s_r)} e^{s_r t} \tag{2.11}$$

In general, the roots s_r of $D(s)$ are complex. For physical systems, the a's in $D(s)$, Eq. (2.8), are real; then the s_r's have complex conjugate pairs. But if all s_r's have negative real parts, then $y_c(t)$ will decrease exponentially with respect to time, and eventually $y_c(t) \to 0$. Then the system is stable.

If the forcing function $x(t)$ is *sinusoidal*, it can be written as

$$x(t) = x_m e^{i\omega t} \tag{2.12}$$

where x_m is the amplitude and ω is the circular frequency. Then, according to the dictionary,

$$X(s) = x_m \frac{1}{s - i\omega}$$

Therefore the second term of Eq. (2.10) is now

$$\frac{x_m}{(s - i\omega)D(s)}$$

We can generalize this to include systems determined by a set of simultaneous equations by putting another polynomial $N(s)$ of order lower than n in the numerator. Then the Laplace transform $Y_i(s)$ of the particular integral is

$$Y_i(s) = F(s)X(s) = \frac{N(s)}{D(s)} X(s) = \frac{x_m N(s)}{(s - i\omega)D(s)} \tag{2.13}$$

When $N(s) \equiv 1$, the problem is reduced to the simpler one specified by Eq. (2.10). Now the partial-fraction rule can be applied again. But the polynomial in the denominator is now $(s - i\omega)D(s)$, and the roots are $s_1, s_2, \ldots, s_n$ and $i\omega$. Thus

$$Y_i(s) = \left[\frac{N(i\omega)}{D(i\omega)}\frac{1}{(s - i\omega)} + \sum_{r=1}^{n} \frac{N(s_r)}{(s_r - i\omega)D'(s_r)}\frac{1}{(s - s_r)}\right] x_m \quad (2.14)$$

Therefore the particular integral $y_i(t)$ due to a sinusoidal forcing function of form (2.12) is

$$y_i(t) = x_m \left[\frac{N(i\omega)}{D(i\omega)} e^{i\omega t} + \sum_{r=1}^{n} \frac{N(s_r)}{(s_r - i\omega)D'(s_r)} e^{s_r t}\right] \quad (2.15)$$

For stable systems, all s_r's have negative real parts. Thus the second part of $y_i(t)$ vanishes as $t \to \infty$. The remaining part is the steady solution, and the ratio of steady solution to the forcing function is thus simply given by

$$\frac{[y_i(t)]_{\text{steady}}}{x(t)} = \frac{N(i\omega)}{D(i\omega)} = F(i\omega) \quad (2.16)$$

This equation gives a very direct way of computing the steady-state solution under a sinusoidal forcing function.

When the frequency ω of the forcing function decreases to zero, the forcing function is reduced to a constant, nonvarying with respect to time. Equation (2.16) then indicates that $F(0)$ is the ratio of y to x when x is a constant. This is the physical meaning of the value of $F(s)$ at $s = 0$. We shall use this interpretation frequently in our discussions.

2.5 Response to Unit Impulse. The forcing function $x(t)$ need not be a continuous function. It may be a unit impulse applied at the time instant $t = 0$, *i.e.*,

$$x(t) = 0 \qquad \text{for } t \neq 0$$
$$x(t) \to \infty \qquad \text{for } t = 0$$

and

$$\int_0^\infty x(t)\, dt = 1$$

Then the Laplace transform $X(s)$ of the forcing function is simply equal to 1. The Laplace transform of the response to the unit impulse of the generalized system as described by Eq. (2.13) is simply

$$Y_i(s) = \frac{N(s)}{D(s)} \cdot 1 = F(s) \quad (2.17)$$

The solution $y(t)$ due to this unit impulse is usually denoted as $h(t)$. According to our inversion formula, Eq. (2.2),

$$h(t) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{st} F(s)\, ds \quad (2.18)$$

When the system is stable, the roots s_r all have negative real parts. Then the singularities of $F(s)$ all lie to the left of the imaginary axis of the complex s plane. Then the imaginary axis can be taken as the path of integration for the $h(t)$, that is, γ in Eq. (2.18) can be set equal to zero.

CHAPTER 3

INPUT, OUTPUT, AND TRANSFER FUNCTION

We have seen in the previous chapter that by the application of the Laplace transform, the behavior of a linear system with constant coefficients is made to depend essentially on the polynomial $D(s)$, given by Eq. (2.8), formed out of the coefficients of the differential equation. Even in the generalized case, if the initial values of $y(t)$ and the initial derivatives necessary to specify the problem are all zero, the behavior of the system is completely determined by the ratio $N(s)/D(s)$ of two polynomials. This ratio is denoted as $F(s)$. If the Laplace transform of the forcing function is $X(s)$ and the Laplace transform of the particular integral is $Y_i(s)$, Eq. (2.13) gives

$$Y_i(s) = F(s)X(s) \tag{3.1}$$

This equation can be considered as an operator equation: $X(s)$ when operated by $F(s)$ gives $Y_i(s)$, or $F(s)$ transfers $X(s)$ into $Y_i(s)$. Therefore $F(s)$ is called the *transfer function.* $X(s)$ is the Laplace transform of the *input* $x(t)$, and $Y_i(s)$ is the Laplace transform of the *output* $y_i(t)$. In order to specify the fact that $y_i(t)$ implies the particular integral only, without the complementary function introduced by the initial conditions, $y_i(t)$ is called the *output due to input.* Then the complementary function $y_c(t)$ is called the *output due to initial conditions.*

The advantage of the Laplace transform method is thus to reduce a problem in differential equations to one of algebraic operation. The step of going from $Y(s)$ to $y(t)$ is seldom necessary, because the behavior of $y(t)$ is fully determined by $Y(s)$. Thus it is possible to translate the engineering requirements on $y(t)$ to a set of requirements on $Y(s)$ or, with the input characteristics specified, to a set of requirements on $F(s)$, the transfer function. The study and design of a system by means of the transfer function are the fundamental technique in servomechanism engineering. In this chapter, we shall expound this technique by a series of examples.

3.1 First-order Systems. As a first example, consider a cantilever spring which has a dashpot at one end and a sliding movement on the other end (Fig. 3.1). The position of the end with the dashpot is denoted

by $y(t)$, the position of the sliding end by $x(t)$. Because of the dashpot, $y(t)$ is not equal to $x(t)$ but lags behind $x(t)$. The problem is to study the output $y(t)$ when the sliding end is made to describe a specified motion. $x(t)$ is thus the input.

Let k be the spring constant and c be the damping coefficient of the dashpot. Then if the motion is slow enough for the inertial forces to be neglected, the equilibrium of forces requires

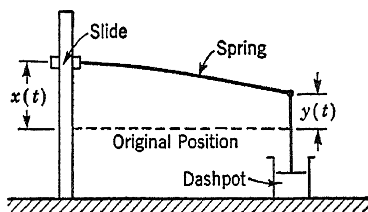


FIG. 3.1

$$c\frac{dy}{dt} + k(y - x) = 0$$

If we specify a characteristic time τ_1 as

$$\tau_1 = \frac{c}{k} \tag{3.2}$$

Then the equation of motion can be written as

$$\tau_1 \frac{dy}{dt} + y = x \tag{3.3}$$

The initial condition is simply

$$y(0) = y_0 \tag{3.4}$$

By multiplying Eq. (3.3) by e^{-st} and integrating from $t = 0$ to $t = \infty$, we have the transformed equation

$$(\tau_1 s + 1)Y(s) = X(s) + \tau_1 y_0$$

Therefore

$$Y(s) = \frac{X(s)}{\tau_1 s + 1} + \frac{\tau_1 y_0}{\tau_1 s + 1} \tag{3.5}$$

Hence the output due to input is given by

$$Y_i(s) = \frac{1}{\tau_1 s + 1} X(s) \tag{3.6}$$

and the output due to the initial condition is given by

$$Y_c(s) = \frac{\tau_1 y_0}{\tau_1 s + 1} \tag{3.7}$$

The transfer function $F(s)$ is thus

$$F(s) = \frac{1}{\tau_1 s + 1} \tag{3.8}$$

Equation (3.6) can be represented graphically as shown in Fig. 3.2. This simple visual aid is very helpful in picturing the situation and is generally called the block diagram.

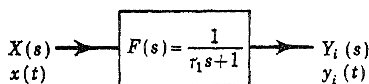


FIG. 3.2

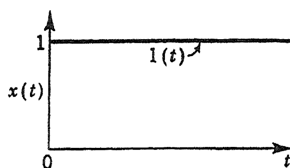


FIG. 3.3

Let us investigate the output for a few special cases of the input $x(t)$. Consider first the case when $x(t)$ is the unit step function $1(t)$ shown in Fig. 3.3. Then

$$X(s) = \int_0^\infty e^{-st}\,dt = \frac{1}{s}$$

and

$$Y_i(s) = \frac{1}{s(\tau_1 s + 1)} = \frac{1}{s} - \frac{1}{s + (1/\tau_1)}$$

Thus the output due to input is, according to our dictionary, Table 2.1,

$$y_i(t) = (1 - e^{-t/\tau_1}) \tag{3.9}$$

The output due to the initial condition is, Eq. (3.7),

$$y_c(t) = y_0 e^{-t/\tau_1} \tag{3.10}$$

These output characteristics are shown in Fig. 3.4. Thus the output due to the initial condition is a pure subsidence with the characteristic time τ_1. The output due to input is an exponential approach to the asymptote, again with the characteristic time τ_1. In fact at $t = \tau_1$, the output $y_i(t)$ reaches 63 per cent of the asymptotic final value.

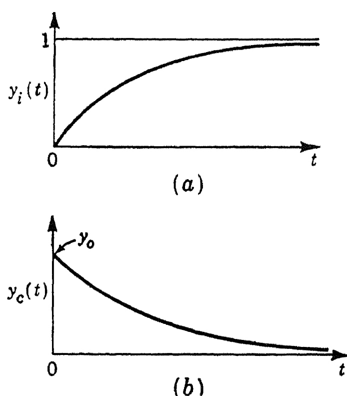


FIG. 3.4

The error signal $e(t)$, defined as the difference between the input $x(t)$ and the output $y_i(t)$ is, for the case under study,

$$e(t) = x(t) - y_i(t) = e^{-t/\tau_1} \tag{3.11}$$

Therefore the error signal vanishes as $t \to \infty$.

Consider now another example of the input. Let the input be sinusoidal, or

$$x(t) = x_m e^{i\omega t}$$

where x_m is the amplitude and ω is the frequency. Then

$$X(s) = \frac{x_m}{s - i\omega} \tag{3.12}$$

The output due to the initial condition is the same as before. The output due to input is given by

$$Y_i(s) = x_m \frac{1}{(s - i\omega)(\tau_1 s + 1)} = \frac{x_m}{1 + i\omega\tau_1}\left(-\frac{1}{s + (1/\tau_1)} + \frac{1}{s - i\omega}\right)$$

Therefore, according to our dictionary, the output $y_i(t)$ is

$$y_i(t) = -\frac{x_m}{1 + i\omega\tau_1} e^{-t/\tau_1} + \frac{x_m}{1 + i\omega\tau_1} e^{i\omega t}$$

The first term is a pure subsidence, and the second term is the steady-state output. Thus

$$\frac{[y(t)]_{\text{steady}}}{x(t)} = \frac{1}{1 + i\omega\tau_1} = F(i\omega)$$

This is in full agreement with our general result given in Eq. (2.16). Since

$$\frac{1}{1 + i\omega\tau_1} = \frac{1}{\sqrt{1 + \omega^2\tau_1^2}} e^{-i\tan^{-1}\omega\tau_1} \tag{3.13}$$

the steady-state output can be expressed as

$$[y(t)]_{\text{steady}} = \frac{x_m}{\sqrt{1 + \omega^2\tau_1^2}} e^{i(\omega t - \tan^{-1}\omega\tau_1)}$$

Therefore the amplitude of the steady-state output is reduced by the factor $1/\sqrt{1 + \omega^2\tau_1^2}$ in comparison with the input, and the phase of the output lags behind the input by the amount $\tan^{-1}\omega\tau_1$. For low-frequency inputs,

$$[y(t)]_{\text{steady}} \approx x_m e^{i\omega(t - \tau_1)} \qquad \tau_1\omega \ll 1 \tag{3.14}$$

That is, the amplitude is not modified, but there is a time lag equal to the characteristic time τ_1 of the transfer function. For high-frequency inputs,

$$[y(t)]_{\text{steady}} \approx \frac{x_m}{\omega\tau_1} e^{i[\omega t - (\pi/2)]} \qquad \tau_1\omega \gg 1 \tag{3.15}$$

Then the amplitude is reduced by the factor $1/\omega\tau_1$, and the phase lags by $\pi/2$. These characteristics of the output are shown in Fig. 3.5.

3.2 Representations of the Transfer Function. The transfer function $F(s)$ is a function of the complex variable s. Since it is generally the ratio of two polynomials in s, the function $F(s)$ is determined up to a

constant by the zeros and the poles of $F(s)$. This constant can be fixed by knowing the value of $F(s)$ at any particular s. The most convenient s is the origin. In fact,

$$|F(0)| = K \tag{3.16}$$

has a physical meaning: it is the ratio of output to input with a constant, non-time-varying input. K is actually called the *gain* of the system. Therefore the transfer function $F(s)$ is uniquely determined by the gain, the zeros, and the poles. This is one possible representation of the transfer function. For example, the simple transfer function specified by Eq. (3.8) has a gain of unity, a simple pole at $-1/\tau_1$, and no zero.

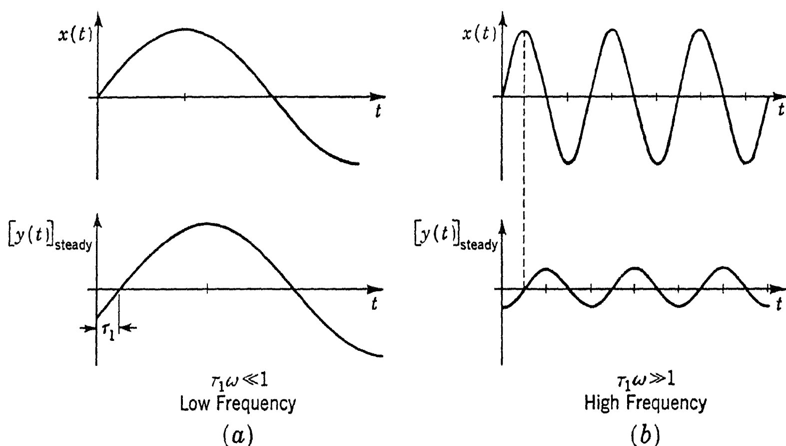


FIG. 3.5

If *both the real and the imaginary parts* of $F(s)$ along the imaginary axis of the s plane are given, then by the principle of analytical continuation, $F(s)$ is essentially determined for any s. Therefore another possible representation of $F(s)$ is the complex function $F(i\omega)$, where ω is real. For physical systems, the coefficients in the numerator polynomial $N(s)$ and the denominator polynomial $D(s)$ of $F(s)$ are all real. Then if we denote the complex conjugate of F by $\bar{F}$,

$$F(-i\omega) = \overline{F(i\omega)} \tag{3.17}$$

Therefore for physical systems, knowledge of $F(i\omega)$ for $\omega \geq 0$ will be sufficient for the determination of F for any s. But we know from Eq. (2.16) that $F(i\omega)$ is the ratio of steady output to the sinusoidal input of frequency ω. $F(i\omega)$ for all values of ω is called the *frequency response* of the system. Therefore the frequency response is another representation of the transfer function. The frequency response of our simple first-order system is given by Eq. (3.13).

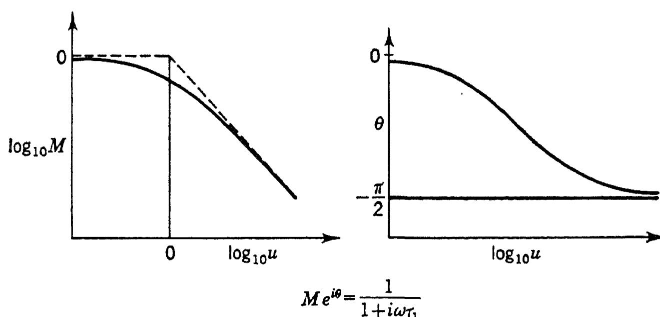
One way to present the frequency response was devised by H. W. Bode and is called the *Bode diagram*. Let the magnitude of $F(i\omega)$ be M and the argument be θ, *i.e.*,

$$F(i\omega) = Me^{i\theta} \tag{3.18}$$

The Bode diagram consists of the plots of $\log M$ and θ against $\log \omega$. The choice of a logarithmic scale for M but not for θ will be made meaningful by later discussions. For our simple system of Eq. (3.13),

$$\left.\begin{aligned} M &= \frac{1}{\sqrt{1 + \omega^2\tau_1^2}} = \frac{1}{\sqrt{1 + u^2}} \\ \text{and} \qquad \theta &= -\tan^{-1}\omega\tau_1 = -\tan^{-1} u \end{aligned}\right\} u = \omega\tau_1 \tag{3.19}$$

where u is the dimensionless frequency. The Bode diagram of this system is shown in Fig. 3.6. The behavior of the frequency response



$$Me^{i\theta} = \frac{1}{1 + i\omega\tau_1}$$

FIG. 3.6

at low and at high frequencies is that already indicated by Eqs. (3.14) and (3.15). As $u \to \infty$, the graph of $\log_{10} M$ against $\log_{10} u$ has the slope -1. For small values of u, the slope is nearly 0. Therefore the $M \sim u$ diagram for a first-order system can be approximated by two straight lines.

In acoustical and electrical literature, it is customary to plot $20 \log_{10} M$ instead of $\log_{10} M$ in order to convert the amplitude units into *decibels*. A doubling in frequency is called an octave, and thus the region of the plot in Fig. 3.6 where the $\log_{10} M$ curve has a slope of -1 is described as a region of slope $-20 \log_{10} 2 = -6.02$ db per octave. We note also that on the same plot, the approximate $\log_{10} M$ line goes through 0 at $u = 1$, that is, at $\omega = 1/\tau_1$. Therefore we can measure the frequency response of a first-order system and plot the measurement as indicated. The characteristic time τ_1 of the system can easily be estimated by noting the frequency at which a straight-line approximation for large values of ω crosses the horizontal axis.

Another way to present the frequency response was devised by H.

Nyquist and is called the *Nyquist diagram*. Here the complex quantity $F(i\omega)$ or $1/F(i\omega)$ is directly plotted in the complex F or $1/F$ plane. The parameter of the curve is the frequency ω. For a simple first-order system, $F(i\omega)$ is a semicircle, starting at 1 for $\omega = 0$, going through $1/(1 + i) = (1/\sqrt{2})(1 - i)$ at $\omega\tau_1 = u = 1$, and ending at the origin for $\omega \to \infty$. The $1/F$ diagram is much simpler: $1/F = 1 + i\omega\tau_1$, and

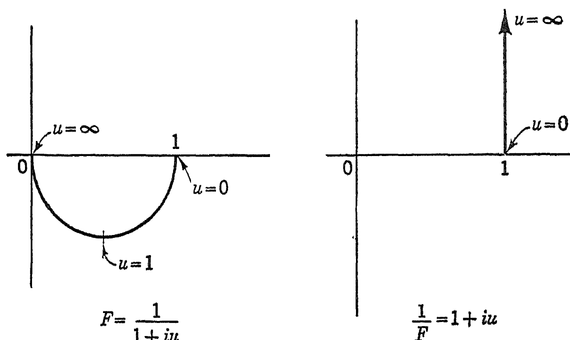


Fig. 3.7

thus the diagram is simply a straight line parallel to the imaginary axis. These Nyquist diagrams for the first-order system are shown in Fig. 3.7.

3.3 Examples of First-order Systems. There are many elements of a complex system that can be approximated by a first-order transfer function. We shall briefly discuss a number of examples of such elements in this section, together with the proper diagrams for their frequency responses.

Integrator. An electric motor whose speed $d\phi/dt$ is proportional to an input voltage v follows the equation

$$\frac{d\phi}{dt} = Kv \tag{3.20}$$

where K is a scale factor. Thus the angular position ϕ of the rotor of the motor is proportional to

$$\int_0^t v\,dt$$

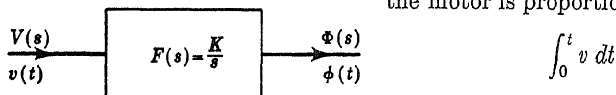


Fig. 3.8

This relation is represented by the block diagram in Fig. 3.8 with the Laplace transforms $V(s)$ and $\Phi(s)$ for v and ϕ. This transfer function $F(s) = K/s$ is the limiting case of the function $1/(\tau_1 s + 1)$ when $\tau_1 \to \infty$, and it is represented by a simple pole at the origin. In order to consider the constant K as the gain of the transfer function $F(s)$, we have to modify the definition first introduced in the previous section. The definition there given is suitable for transfer functions having no zero or

pole at the origin. The gain K of an integrating system, *i.e.*, a system whose transfer function $F(s)$ has a simple pole at the origin $s = 0$, should be defined as

$$K = \lim_{s \to 0} |sF(s)| \tag{3.21}$$

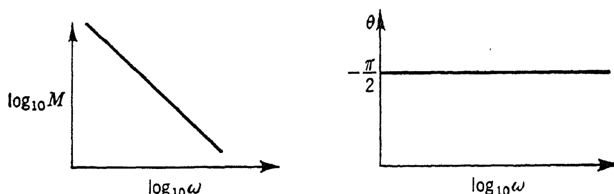
The frequency response is

$$F(i\omega) = \frac{K}{i\omega} = \left(\frac{K}{\omega}\right) e^{-i(\pi/2)}$$

Therefore, according to Eq. (3.18),

$$M = \frac{K}{\omega} \qquad \theta = -\frac{\pi}{2} \tag{3.22}$$

The Bode diagram is thus that shown in Fig. 3.9, and the Nyquist diagram is that shown in Fig. 3.10.



$F(i\omega) = \frac{K}{i\omega} = Me^{i\theta}$

Fig. 3.9

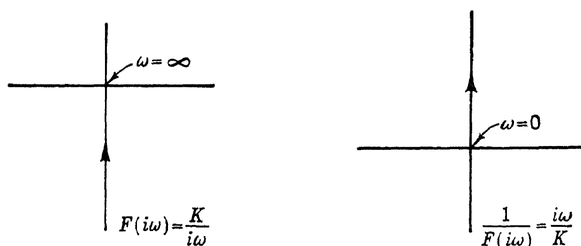


Fig. 3.10

Differentiator. A rate gyro gives a voltage output v proportional to the angular velocity $d\phi/dt$ of the precession axis, *i.e.*,

$$v = K\frac{d\phi}{dt}$$

where K is the factor of proportionality. This case is the inverse of the preceding one. The transfer function $F(s) = Ks$ has a zero at the origin. Thus the gain of a differentiating system, *i.e.*, a system whose transfer function $F(s)$ as a simple zero at the origin $s = 0$, should be

defined as

$$K = \lim_{s\to 0}\left|\frac{F(s)}{s}\right| \tag{3.23}$$

The block diagram is shown in Fig. 3.11, the Bode diagram in Fig. 3.12, and the Nyquist diagram in Fig. 3.13.

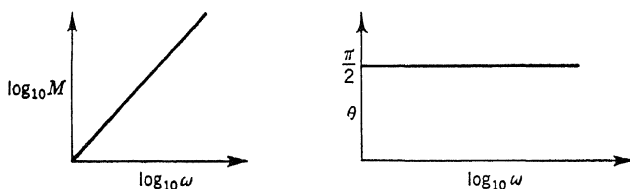
$\Phi(s)$, $\phi(t)$ → $F(s) = Ks$ → $V(s)$, $v(t)$

Fig. 3.11

Simple Lag Network. Consider the resistance and capacitance network of Fig. 3.14. If j is the current flowing in the resistance R and the capacitance C and if there is no charge in the capacitance at $t = 0$, then

$$jR + \frac{1}{C}\int_0^t j(t)\,dt = v_1$$
$$\frac{1}{C}\int_0^t j(t)\,dt = v_2$$

By multiplying these equations by e^{-st} and integrating from $t = 0$ to



$F(i\omega) = Ki\omega = Me^{i\theta}$

Fig. 3.12

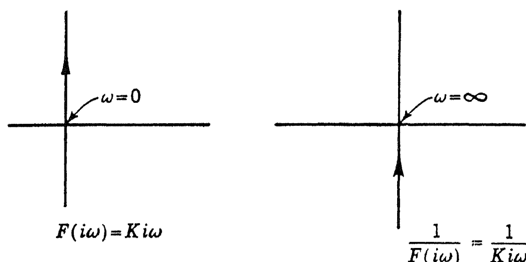


Fig. 3.13

$t = \infty$, we have

$$\left(R + \frac{1}{Cs}\right)J(s) = V_1(s)$$
$$\frac{1}{Cs}J(s) = V_2(s)$$

Therefore

$$\frac{V_2(s)}{V_1(s)} = F(s) = \frac{1}{1 + RCs} \tag{3.24}$$

Hence the transfer function of this RC circuit is the same as the cantilever spring with dashpot, and the characteristic time is $\tau_1 = RC$. The Bode diagram and the Nyquist diagram are thus given by Figs. 3.6 and 3.7. This circuit is frequently used in order to introduce a phase lag into a system.

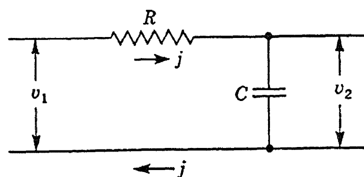


Fig. 3.14

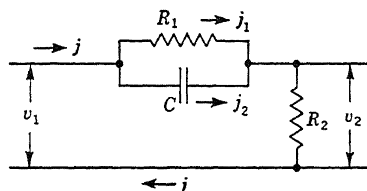


Fig. 3.15

Lead Network. A more complex circuit is that shown in Fig. 3.15. The controlling equations are

$$j = j_1 + j_2$$

$$R_1 j_1 = \frac{1}{C} \int_0^t j_2(t)\, dt$$

and

$$v_1 = R_1 j_1 + R_2 j$$

$$v_2 = R_2 j$$

The corresponding transformed equations are

$$J = J_1 + J_2$$

$$R_1 J_1 = \frac{1}{Cs} J_2$$

and

$$V_1 = R_1 J_1 + R_2 J$$

$$V_2 = R_2 J$$

Therefore

$$\frac{V_2(s)}{V_1(s)} = F(s) = \frac{R_2 + R_1 R_2 Cs}{(R_1 + R_2) + R_1 R_2 Cs}$$

Hence the gain is

$$K = \frac{R_2}{R_1 + R_2} = r \tag{3.25}$$

and it is necessarily less than unity and is generally between 0.1 and 1. If we introduce the symbol ω_1 as

$$\omega_1 = \frac{R_1 + R_2}{R_1 R_2 C} \tag{3.26}$$

then the transfer function can be written as

$$F(s) = r \frac{1 + (s/r\omega_1)}{1 + (s/\omega_1)} \tag{3.27}$$

Therefore the transfer function has a zero at $-r\omega_1$ and a pole at $-\omega_1$.

The frequency response is then

$$F(i\omega) = \frac{r\omega_1 + i\omega}{\omega_1 + i\omega} \tag{3.28}$$

If we introduce the nondimensional frequency u as

$$u = \frac{1}{\sqrt{r}} \frac{\omega}{\omega_1} \tag{3.29}$$

then

$$M = \sqrt{r} \sqrt{\frac{1 + (u^2/r)}{(1/r) + u^2}} \qquad \theta = \tan^{-1} \frac{u}{\sqrt{r}} - \tan^{-1} (\sqrt{r}\, u) \tag{3.30}$$

Hence

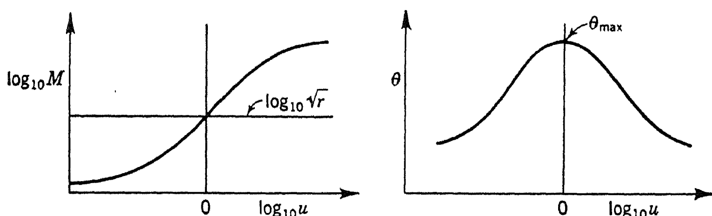
$$\log_{10} M(u) = \log_{10} \sqrt{r} + \log_{10} \sqrt{\frac{1 + (u^2/r)}{(1/r) + u^2}}$$

$$= \log_{10} \sqrt{r} - \log_{10} \sqrt{\frac{1 + (1/ru^2)}{(1/r) + (1/u^2)}}$$

and

$$\theta(u) = \theta\left(\frac{1}{u}\right)$$

The Bode diagram has thus a certain symmetry with respect to $u = 1$,



$$F(i\omega) = \frac{r\omega_1 + i\omega}{\omega_1 + i\omega}, \; u = \frac{1}{\sqrt{r}} \frac{\omega}{\omega_1}$$

Fig. 3.16

as shown in Fig. 3.16. The maximum value of θ occurs at $u = 1$ and is equal to

$$\theta_{\max} = \tan^{-1} \frac{1}{\sqrt{r}} - \tan^{-1} \sqrt{r} = \frac{\pi}{2} - 2 \tan^{-1} \sqrt{r} \tag{3.31}$$

Therefore this circuit gives a considerable phase lead over a band of frequencies. For very large ω, $M = 1$. For very small ω, $M = r$.

Restricted Lag Network. The *RC* circuit shown in Fig. 3.17 has the following transfer function:

$$\frac{V_2(s)}{V_1(s)} = F(s) = \frac{1 + R_1 Cs}{1 + (R_1 + R)Cs}$$

The gain of the system is thus unity. If we introduce the parameters ω_1 and r defined as

$$\omega_1 = \frac{1}{R_1 C} \qquad r = \frac{R_1}{R + R_1} \tag{3.32}$$

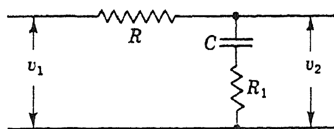


Fig. 3.17

then the transfer function is

$$F(s) = \frac{1 + (s/\omega_1)}{1 + (s/r\omega_1)} \tag{3.33}$$

By comparing this equation with Eq. (3.28) for the lead network, we see that the two circuits have transfer functions that are reciprocal to each other. In fact the frequency response in the present case can be written as

$$F(i\omega) = \frac{1 + i\,(\omega/\omega_1)}{1 + i\,(\omega/r\omega_1)} = \frac{1 + i\sqrt{r}\,u}{1 + i\,(1/\sqrt{r})\,u}$$

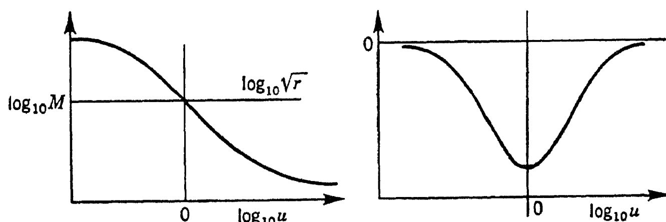
where u, the dimensionless frequency, is

$$u = \frac{1}{\sqrt{r}}\frac{\omega}{\omega_1} \tag{3.34}$$

Thus

$$M = \sqrt{r}\sqrt{\frac{(1/r) + u^2}{1 + (u^2/r)}} \qquad \theta = \tan^{-1}(\sqrt{r}\,u) - \tan^{-1}\frac{u}{\sqrt{r}} \tag{3.35}$$

The corresponding Bode diagram is thus that shown in Fig. 3.18. There is a phase lag for a range of the frequency. The maximum phase lag



$$F(i\omega) = \frac{1 + i\frac{\omega}{\omega_1}}{1 + i\frac{\omega}{r\omega_1}}, \quad u = \frac{1}{\sqrt{r}}\frac{\omega}{\omega_1}$$

Fig. 3.18

occurs at $u = 1$, or $\omega = \sqrt{r}\,\omega_1$, and its magnitude is given by Eq. (3.31).

Simplified Rolling Motion of an Airplane. Let I be the moment of inertia of the airplane about its longitudinal axis, ϕ the roll angle, L_p the aerodynamic damping due to roll, and $k\delta$ the torque applied by the aileron deflection δ. The equation for the roll angle ϕ is thus

$$I\frac{d^2\phi}{dt^2} + L_p\frac{d\phi}{dt} = k\delta$$

Now let $p = d\phi/dt$ be the roll speed; then the above equation becomes

$$I\frac{dp}{dt} + L_p p = k\delta$$

If the roll speed is zero at $t = 0$, then the transformed equation is

$$(Is + L_p)P(s) = k\,\Delta(s)$$

The transfer function $F(s)$ is thus

$$\frac{P(s)}{\Delta(s)} = F(s) = \frac{k}{Is + L_p} = \frac{k}{L_p}\frac{1}{1 + (I/L_p)s} \tag{3.36}$$

The behavior of the system, as determined by the transfer function, is thus similar to the cantilever spring with dashpot and the simple lag network. Here the characteristic time τ_1 is I/L_p. If the damping is very small, then $\tau_1 \to \infty$ and the behavior of the system becomes that of the simple integrator.

3.4 Second-order Systems. Let us return to the cantilever spring with a dashpot (Fig. 3.1). But now we attach a mass m to the dashpot end. The mass will introduce an inertial force $m\,d^2y/dt^2$, and the equation of motion is now

$$m\frac{d^2y}{dt^2} + c\frac{dy}{dt} + ky = kx$$

with the initial conditions

$$\left.\begin{aligned} y(0) &= y_0 \\ \left(\frac{dy}{dt}\right)_{t=0} &= y_0^{(1)} \end{aligned}\right\} \tag{3.37}$$

The differential equation of motion can be rewritten in a more convenient form by introducing the following parameters:

$$\begin{aligned} \omega_0^2 &= \frac{k}{m} \\ \zeta &= \frac{c/m}{2\omega_0} \end{aligned} \tag{3.38}$$

ω_0 is thus the natural frequency of the mass-spring combination when the dashpot is absent. ζ is the ratio of actual damping to critical damping. The meaning of this nondimensional parameter will be made clear presently. Then the differential equation becomes

$$\frac{d^2y}{dt^2} + 2\zeta\omega_0\frac{dy}{dt} + \omega_0^2 y = \omega_0^2 x \tag{3.39}$$

Equation (3.39) together with the initial conditions of Eq. (3.37) can be converted into the following relation in terms of Laplace transforms:

$$(s^2 + 2\zeta\omega_0 s + \omega_0^2)Y(s) = \omega_0^2 X(s) + y_0^{(1)} + (s + 2\zeta\omega_0)y_0$$

The output due to initial conditions is then given by

$$Y_c(s) = \frac{y_0 s + (y_0^{(1)} + 2\zeta\omega_0 y_0)}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \tag{3.40}$$

The transfer function is then

$$F(s) = \frac{Y_i(s)}{X(s)} = \frac{1}{(s/\omega_0)^2 + 2\zeta(s/\omega_0) + 1} \tag{3.41}$$

The transfer function thus has a gain $K = 1$ and no zeros. It has two simple poles at

$$\left.\begin{aligned} \frac{s_1}{\omega_0} &= -\zeta + \sqrt{\zeta^2 - 1} \\ \frac{s_2}{\omega_0} &= -\zeta - \sqrt{\zeta^2 - 1} \end{aligned}\right\} \zeta^2 > 1 \tag{3.42}$$

When the damping coefficient of the dashpot is smaller than the critical damping, the value of ζ will be less than unity. In that case, the poles s_1 and s_2 will be complex conjugates, having real and imaginary parts λ and ν, respectively:

$$\left.\begin{aligned} s_1/\omega_0 &= -\zeta + i\sqrt{1 - \zeta^2} = (\lambda + i\nu)/\omega_0 = e^{i\varphi_1} \\ s_2/\omega_0 &= -\zeta - i\sqrt{1 - \zeta^2} = (\lambda - i\nu)/\omega_0 = e^{-i\varphi_1} \end{aligned}\right\} \zeta^2 < 1 \tag{3.43}$$

where the last expression is possible because the absolute value of either s_1/ω_0 or s_2/ω_0 is one. For positive damping, λ is a negative number.

The output $y_c(t)$ due to initial conditions can be easily determined from Eq. (3.40). Thus, for $\zeta^2 < 1$, the poles of the transfer function are given by Eq. (3.43), and we have

$$y_c(t) = \frac{y_0^{(1)}}{\nu} e^{\lambda t} \sin \nu t + y_0 e^{\lambda t} \cos \nu t + \frac{-\lambda}{\nu} y_0 e^{\lambda t} \sin \nu t \tag{3.44}$$

Since λ is a negative number, the output is damped, but nevertheless it is a damped sinusoidal function. If, on the other hand, $\zeta^2 > 1$, then the output is a pure subsidence. Thus for dampings greater than the critical value, there is no oscillation in the output $y_c(t)$. This is the meaning of critical damping.

Now let us assume that the input $x(t)$ is the unit step function $1(t)$ shown in Fig. 3.3. Then $X(s) = 1/s$, and for $\zeta^2 < 1$,

$$Y_i(s) = \frac{\omega_0^2}{s[(s - \lambda)^2 + \nu^2]}$$

The output $y_i(t)$ due to input is then

$$y_i(t) = 1 - \left[\cos \nu t + \left(\frac{-\lambda}{\nu}\right) \sin \nu t\right] e^{\lambda t} \tag{3.45}$$

When $\zeta^2 > 1$, the output is not oscillatory and is calculated as

$$y_i(t) = 1 - \frac{1}{2\sqrt{\zeta^2 - 1}}\left[\frac{e^{s_1 t}}{\zeta - \sqrt{\zeta^2 - 1}} - \frac{e^{s_2 t}}{\zeta + \sqrt{\zeta^2 - 1}}\right] \tag{3.46}$$

where s_1 and s_2 are given by Eq. (3.42). Such behaviors of the output $y_i(t)$ are shown in Fig. 3.19 for various values of the damping ratio ζ. It is seen that for a quick approach to the asymptotic value, ζ should not be too large. On the other hand, if ζ is too small, there will be rather

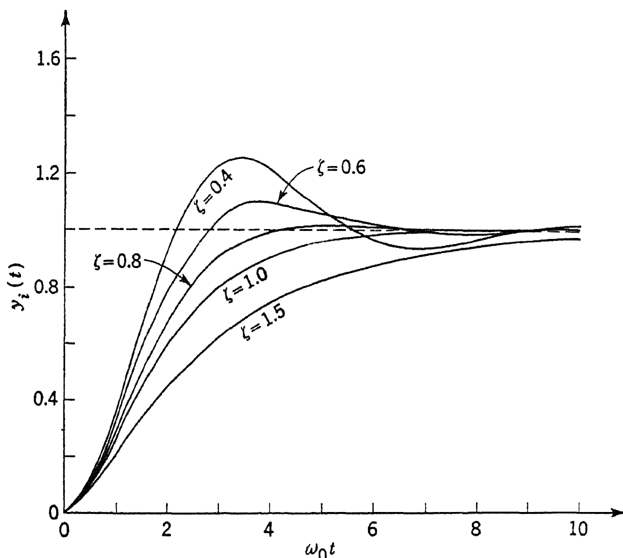


FIG. 3.19

persistent oscillations with high peaks. Here an engineering compromise has to be made, and the usual practice is to select a damping ratio ζ between 0.4 and 1.

If the input is a sinusoidal oscillation with amplitude x_m and frequency ω as specified by Eq. (3.11), then

$$Y_i(s) = \frac{x_m}{s - i\omega} F(s) = \frac{x_m}{s - i\omega} \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

Therefore the output $y_i(t)$ due to input is, for $\zeta^2 < 1$,

$$y_i(t) = x_m F(i\omega) e^{i\omega t} + \frac{x_m}{2i\nu} \frac{\omega_0^2}{\lambda + i(\nu - \omega)} e^{(\lambda + i\nu)t} - \frac{x_m}{2i\nu} \frac{\omega_0^2}{\lambda - i(\nu - \omega)} e^{(\lambda - i\nu)t} \tag{3.47}$$

where λ and ν are given by Eq. (3.43). Since λ is negative for positive damping, the steady-state output is again simply given by the first

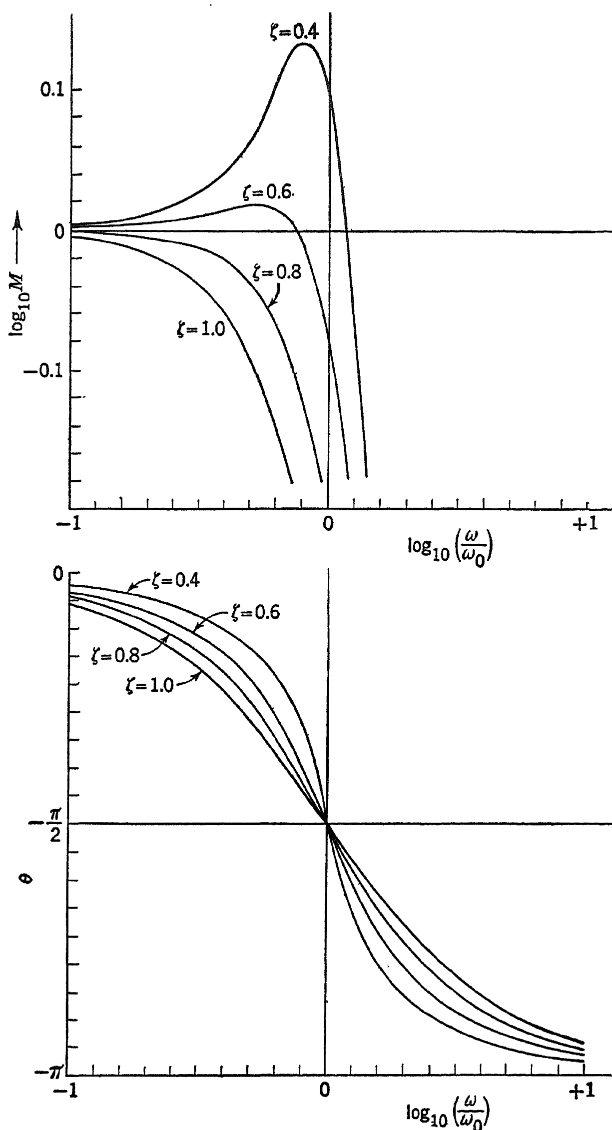


FIG. 3.20

term of Eq. (3.47). This is in agreement with our general result of Eq. (2.16).

The frequency response of our second-order system is, according to Eq. (3.41),

$$F(i\omega) = Me^{i\theta} = \frac{1}{[1 - (\omega/\omega_0)^2] + 2i\zeta(\omega/\omega_0)}$$

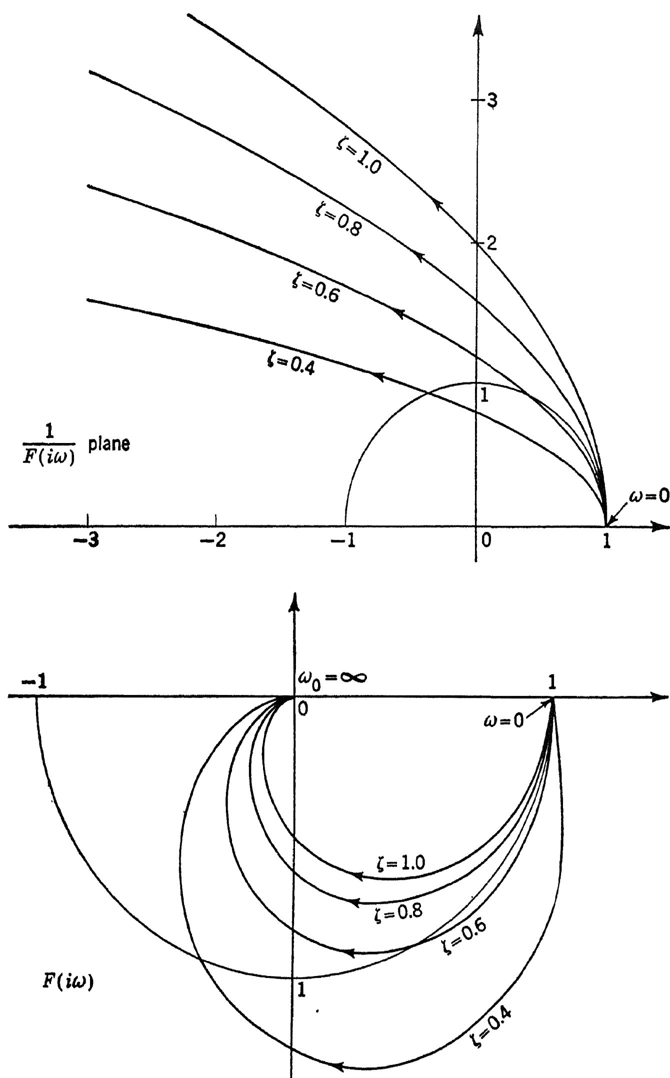


Fig. 3.21

Therefore

$$M = \frac{1}{\sqrt{[1 - (\omega/\omega_0)^2]^2 + [2\zeta(\omega/\omega_0)]^2}} \tag{3.48}$$

and

$$\tan\theta = -\frac{2\zeta(\omega/\omega_0)}{1 - (\omega/\omega_0)^2}$$

The corresponding Bode diagram is shown in Fig. 3.20. The maxima of M occur near $\omega/\omega_0 = 1$, where $M \approx \frac{1}{2\zeta}$ and $\theta \approx -\pi/2$. As $\omega/\omega_0 \to \infty$,

$\theta \to -\pi$ and $M \sim 1/(\omega/\omega_0)^2$, or $\log M \sim -2 \log (\omega/\omega_0)$. The acoustical engineer will then say that the slope for high frequency is -12.04 db per octave.

The Nyquist diagram for our second-order system is shown in Fig. 3.21.

Other physical systems can usually be approximated by a second-order transfer function. The hydraulic servo system is one example. A better approximation to the behavior of the rate gyro discussed in Sec. 3.3 is the transfer function

$$F(s) = \frac{Ks}{(s/\omega_0)^2 + 2\zeta(s/\omega_0) + 1}$$

This more accurate transfer function should be compared with that given in Fig. 3.11. The transfer function for an accelerometer is

$$F(s) = \frac{Ks^2}{(s/\omega_0)^2 + 2\zeta(s/\omega_0) + 1}$$

The electric motor used as an integrator has a more accurate transfer function

$$F(s) = \frac{K}{s(\tau_1 s + 1)}$$

This transfer function should be compared with the previous crude approximation given in Fig. 3.8. All these transfer functions have a second-order polynomial in the denominator. The constants in them have meanings similar to those in the examples discussed previously.

3.5 Determination of Frequency Response. In the discussions of the previous sections, we have considered the problem of knowing the structural details of a system and of calculating the transfer function $F(s)$ and the frequency response $F(i\omega)$ by elementary physical laws. This procedure of determining the frequency response is thus theoretical, and its accuracy depends upon the accuracy of our knowledge of the system. Very often in engineering practice, our knowledge of the detailed structure of the system is incomplete, or if sufficiently complete, the system is so complicated as to make the theoretical calculation of the frequency response too lengthy to be practical. In such cases, it is often necessary to determine the frequency response experimentally. The simplest method conceptually is to utilize the fact that the ratio of steady output with a sinusoidal input of frequency ω to the input is equal to the frequency response $F(i\omega)$ as shown by Eq. (2.16). The ratio of the amplitudes of output and input is M, and the phase difference between output and input is θ. Therefore this experimental method involves the determination of amplitude ratios and phase differences for a number of frequencies ω in the desired range. It has been applied to problems varying from

such a relatively simple system as a fuel pump[1] to as complicated a system as the longitudinal motion of a complete airplane.[2] The drawback of this method is the lengthy experimentation generally required for a wide range of frequencies. Sometimes it is also difficult to determine the phase difference between the output and the input.

A more efficient method is to excite all frequencies simultaneously instead of individually. The best method for doing this is to use a

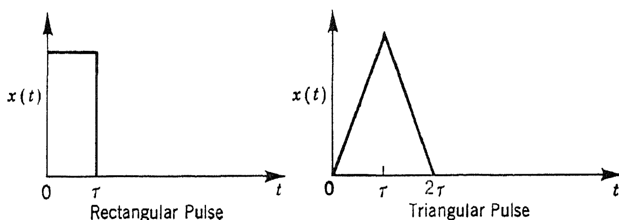


FIG. 3.22

unit impulse as the input. Then, according to Eq. (2.18), for stable systems ($\gamma = 0$),

$$h(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(i\omega)e^{i\omega t}\,d\omega$$
$$= \frac{1}{\pi}\int_{0}^{\infty} [\mathcal{R}F(i\omega)\cos\omega t - \mathcal{I}F(i\omega)\sin\omega t]\,d\omega \tag{3.49}$$

where $\mathcal{R}$ and $\mathcal{I}$ denote the real and the imaginary parts, respectively. This last step is possible because of Eq. (3.17). Equation (3.49) shows that an input impulse excites all the frequencies of the system uniformly. When the response $h(t)$ of the system to a unit impulse is determined, the frequency response can be computed as

$$F(i\omega) = \int_0^{\infty} h(t)e^{-i\omega t}\,dt \tag{3.50}$$

This integration can be carried out numerically for a number of frequencies.

Practically, however, it is difficult to make the input be an impulse. The more practical inputs are a single rectangular pulse and a single triangular pulse, as shown in Fig. 3.22. Such inputs will not excite all frequencies uniformly. But if we make the length τ of the pulse small, the ideal uniform excitation can be approached. This method of pulse excitation has been applied to determine the frequency response of an airplane by R. C. Seamans and his coworkers.[3] They have also

[1] H. Shames, S. C. Himmel, D. Blivas, *NACA TN* 2109 (1950).
[2] W. F. Milliken, *J. Aeronaut. Sci.*, **14,** 493 (1947).
[3] R. C. Seamans, B. P. Blasingame, G. C. Clementson, *J. Aeronaut. Sci.*, **17,** 22 (1950).

developed an approximate method for computing the $F(i\omega)$ from the measured output $y(t)$. The method of data reduction has been generalized to arbitrary inputs by H. J. Curfman and R. A. Gardiner.[1]

3.6 Composition of a System from Elements. The systems studied in Secs. 3.1, 3.3, and 3.4 are really only elements in the much more complicated system generally found to be necessary in stability and control engineering. Take the example of the rolling motion of an airplane. The signal to move the aileron is usually in the form of an electric current. This signal is the input to an "amplifier" and computer group, which is a rather involved electric circuit and may contain vacuum tubes. The behavior of the amplifier and computer is determined by its transfer function $F_1(s)$. The output of the amplifier and computer is then taken to be the input to the hydraulic servo which moves the aileron. The behavior of the hydraulic servo is specified by the servo transfer function

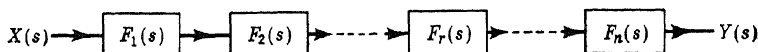


FIG. 3.23

$F_2(s)$. Finally, the output of servo, the aileron motion, is taken as the input to the system representing the airplane dynamics. The airplane dynamics gives the transfer function $F_3(s)$. The output of the airplane dynamics is the rolling motion. Here we have a series connection between the various elements of the system from rolling signal to rolling motion. If the rolling signal is denoted by $x(t)$ and the roll angle by $\phi(t)$, then the Laplace transforms are related by

$$\Phi_i(s) = F_3(s)F_2(s)F_1(s)X(s)$$

Therefore the over-all transfer function of the roll-control system is the product $F_1(s)F_2(s)F_3(s)$. This example also illustrates clearly the fact that a transfer function is generally dimensional: it is the ratio of two quantities of different dimensions. The input, or roll signal, is an electric current; and the output, or roll angle, has the dimension of an angle.

In general, then, if a system is composed of individual elements of transfer functions $F_1(s), F_2(s), \ldots, F_r(s), \ldots, F_n(s)$ with gains $K_1, K_2, \ldots, K_r, \ldots, K_n$ and if the elements are in series (Fig. 3.23), then the over-all transfer function $F(s)$ is the product

$$F(s) = F_1(s)F_2(s) \cdots F_r(s) \cdots F_n(s) \tag{3.51}$$

The gain K of the system is then

$$K = K_1K_2 \cdots K_r \cdots K_n \tag{3.52}$$

From Eq. (3.51), it is evident that the system transfer function $F(s)$ has the totality of the zeros and poles of the individual elements. This fact

[1] H. J. Curfman and R. A. Gardiner, *NACA TR* 984 (1950).

together with the gain K, computed by Eq. (3.52), completely determines the transfer function $F(s)$.

The frequency response of the system is $F(i\omega) = Me^{i\theta}$. If the frequency response of rth element is $M_r e^{i\theta_r}$, then, according to Eq. (3.51),

$$Me^{i\theta} = (M_1 e^{i\theta_1})(M_2 e^{i\theta_2}) \cdots (M_r e^{i\theta_r}) \cdots (M_n e^{i\theta_n})$$

Therefore

$$\left.\begin{aligned} \log_{10} M &= \log_{10} M_1 + \log_{10} M_2 + \cdots + \log_{10} M_r + \cdots \\ &\qquad + \log_{10} M_n \\ \text{and} \quad \theta &= \theta_1 + \theta_2 + \cdots + \theta_r + \cdots \theta_n \end{aligned}\right\} \tag{3.53}$$

Equation (3.53) gives the reason for the choice of a logarithmic scale in the Bode diagram. This choice makes the work of finding the system characteristics easier by requiring only simple addition of the ordinates of individual diagrams.

3.7 Transcendental Transfer Functions. The Laplace-transform method is applicable not only to initial-value problems of linear ordinary differential equations with constant coefficients, but also to linear partial differential equations[1] with coefficients that are independent of the time variable t, and with boundary conditions partially described as initial-value conditions in t. By applying the Laplace transform to the original partial differential equation, the time variable t is removed, and in its place a *parameter* s appears. The resultant equation is a linear differential equation with respect to the remaining independent variables and can be solved as such by using the remaining boundary conditions. The procedure involved here is evidently much more complicated than in the case of ordinary differential equations discussed in Chap. 2. On the other hand, if any two specific quantities in the solution of the transformed equation are compared, they still bear a linear relation. If one of them is considered to be an input and the other an output, the ratio of output to input is still a function of the parameter s and can still be considered as the transfer function $F(s)$. There is, however, one difference: the transfer function is no longer the quotient of two polynomials of s. It is, in general, a transcendental function in s.

For instance, for two-dimensional flows, W. R. Sears[2] has calculated the effects of a small vertical "gust" velocity v in the fluid on an airfoil of chord c in a stream moving horizontally with a uniform velocity U. Let v vary sinusoidally with respect to x, the horizontal coordinate, and

[1] See for instance H. S. Carslaw and J. C. Jaeger, "Operational Methods in Applied Mathematics," Oxford University Press, New York, 1941; or R. V. Churchill, "Modern Operational Methods in Engineering," McGraw-Hill Book Company, Inc., New York, 1944.

[2] W. R. Sears, *J. Aeronaut. Sci.*, **8**, 104 (1941).

t (Fig. 3.24), so that

$$v(x,t) = \alpha_m U e^{i\omega[t-(x/U)]} \tag{3.54}$$

where α_m is the amplitude and ω the "frequency." For this gust velocity, Sears has shown that the lift coefficient C_l, that is, the average lift force per unit area of the airfoil divided by the "dynamic pressure" $\frac{1}{2}\rho U^2$ (ρ is

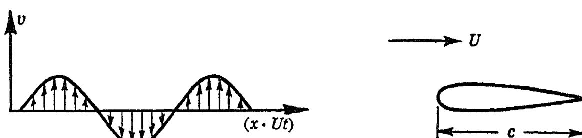


Fig. 3.24

the density of the fluid), is given by

$$C_l = 2\pi \alpha_m e^{i\omega t} \varphi(k) \tag{3.55}$$

where

$$k = \frac{\omega c}{2U} \tag{3.56}$$

and

$$\varphi(k) = \frac{J_0(k)K_1(ik) + iJ_1(k)K_0(ik)}{K_1(ik) + K_0(ik)} \tag{3.57}$$

The J's and K's in Eq. (3.57) are the Bessel functions of the first kind and the modified Bessel functions of the second kind, respectively. Therefore, if we take $X(s)$ as the Laplace transform of $v(0,t)$, the input, and $Y(s)$ as the Laplace transform of $C_l(t)$, the output, then the transfer function $F(s)$ is

$$\frac{Y(s)}{X(s)} = F(s)$$

and the frequency response $F(i\omega)$ is

$$F(i\omega) = \frac{2\pi}{U} \varphi(k) \tag{3.58}$$

The frequency response is thus a transcendental function.

The application of such concepts as the transcendental transfer function and frequency response to the problem of flutter of airplane wings has been demonstrated by J. Dugundji.[1]

[1] J. Dugundji, *J. Aeronaut. Sci.*, **19**, 422 (1952).

CHAPTER 4

FEEDBACK SERVOMECHANISM

In this chapter we shall introduce the central concept of modern stability and control engineering: the concept of feedback. We shall introduce this concept by discussing the simplest systems—linear systems with constant coefficients. We shall show how the feedback can greatly increase the degree of accuracy in control and the rapidity of response to a signal. We shall then explain the principles of designing such feedback servomechanisms for stability and for optimum performance.

4.1 Concept of Feedback. Let us consider the problem of controlling the rotational speed of a turboalternator. The primary purpose here is to keep the speed very close to the normal value. The most elementary approach to this problem would be the so-called *open-cycle* control, where we try to balance the torque generated by the steam turbine and the torque absorbed by the alternator, the load torque. This could be done by measuring the load and by opening the steam throttle accordingly. However, it is to be expected that such balancing cannot be perfect; there is always an error torque, $x(t)$. This error torque tends to accelerate the machine. If we denote by $y(t)$ the speed deviation from the normal, by I the moment of inertia of the rotating parts of the machine, and by c the damping due to the windage loss, then the differential equation is

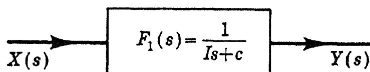


FIG. 4.1

$$I\frac{dy}{dt} + cy = x(t) \qquad (4.1)$$

The block diagram of this system is shown in Fig. 4.1. It is seen that the system is the familiar first-order system studied in the last chapter, that the characteristic time of the system is I/c, and that the ratio of the steady-state value of speed deviation to the error torque is $1/c$. Now I is a very large quantity because of the heavy weight of the rotor of a turboalternator, but c is a very small quantity because of the small windage loss. Hence the characteristic time is extremely long. This means that any speed deviation will persist and be difficult to remove. Furthermore, for small speed deviation, the error torque has to be extremely small, because of the large magnification factor $1/c$. Need-

less to say, this system of keeping the turboalternator at constant speed is quite useless in practice.

Now consider the change in performance caused by changing the system to the so-called *closed-cycle* control. In the closed-cycle control, we make the control torque depend upon the controlled variable. That is, we cause the steam-throttle opening to depend not only on the load but also on the speed deviation y. Let the second component have a factor of proportionality $-k$. When the speed is too high, or $y > 0$, then the throttle is closed, and the accelerating torque is reduced by the amount ky. When the speed is too low, the accelerating torque is increased by the amount ky. Thus the differential equation for y is now

$$I\frac{dy}{dt} + (c + k)y = x(t) \tag{4.2}$$

The only difference between Eqs. (4.1) and (4.2) is the replacement of c by the sum $c + k$. Hence the characteristic time is now $I/(c + k)$, and the ratio of steady-state speed deviation to the error torque is $1/(c + k)$.

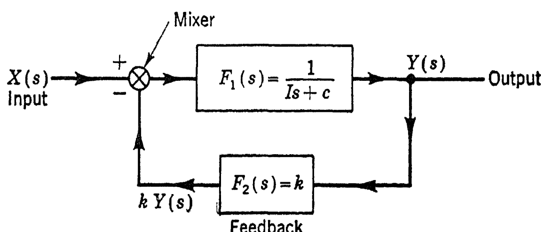


FIG. 4.2

Therefore, in comparison with the open-cycle control, we can greatly reduce both the characteristic time and the speed error by making k very much larger than c. But this can be accomplished quite easily, because c is so small. Therefore the closed-cycle control can be designed for quick response and for accurate control and thus achieve a great improvement in performance.

The block diagram of the closed-cycle control can be drawn as Fig. 4.2, retaining the intrinsic transfer function of the turboalternator shown in Fig. 4.1. In Fig. 4.2, we have introduced a convention in servomechanism engineering: addition or subtraction has to be specifically indicated at the symbol for the *mixer*. If at the junction of two links only a dot is used, no addition or subtraction occurs. The quantity is only "measured." Thus the speed deviation y is measured at the output side and is used to generate the control torque. It is seen from Fig. 4.2 that the closed-cycle control involves a feedback link. The whole system is thus appropriately called a feedback servomechanism.

Although for the simple example analyzed above the advantage of a

feedback servomechanism can be shown by comparing the differential equations (4.1) and (4.2), for more complicated systems the analysis can be conveniently carried out only by the concept of transfer functions. This method is expounded in the following sections.

4.2 Design Criteria of Feedback Servomechanisms. Let us consider a *general feedback servomechanism* with arbitrary transfer functions $F_1(s)$ and $F_2(s)$, similar to that represented by Fig. 4.2. $F_1(s)$ is called the transfer function of the *forward circuit* and $F_2(s)$ the transfer function of the *feedback circuit.* Then the input $X(s)$ and the output $Y(s)$ are related as follows:

$$Y(s) = F_1(s)[X(s) - F_2(s)Y(s)]$$

By solving for $Y(s)$, we have

$$\frac{Y(s)}{X(s)} = \frac{F_1(s)}{1 + F_1(s)F_2(s)} = F_s(s) \tag{4.3}$$

where $F_s(s)$ is thus the system transfer function, or the output-input ratio of the complete system.

It will be convenient for later discussions to indicate explicitly the gains K_1 and K_2 of the transfer functions $F_1(s)$ and $F_2(s)$. Thus we write

$$\left.\begin{aligned} F_1(s) &= K_1G_1(s) \\ F_2(s) &= K_2G_2(s) \end{aligned}\right\} \tag{4.4}$$

It is evident that $G(s)$ is nondimensional; the dimension of the transfer function is all absorbed into the gain K. All information about the "structure" of the transfer function is contained in $G(s)$, that is, $G(s)$ gives the zeros and the poles of the transfer function. In the subsequent discussions, we shall usually think of the effect of the transfer function on the performance of the system as the result of two separate influences: the influence of the locations of the zeros and the poles of the transfer function, *i.e.*, the influence of $G(s)$; and the influence of the magnitude of the gain K. This separation of effects is further justified by the fact that the structure of the transfer function $G(s)$ is controlled by the computer element of the amplifier-computer group in the composite system of $F(s)$. The gain is controlled by the amplifier element of the amplifier-computer group. Moreover, these two controls in the design can be affected almost independently of each other. Therefore $G(s)$ and K can indeed be modified separately and can be considered as separate.

By using Eq. (4.4), Eq. (4.3) can be written as

$$\frac{Y(s)}{X(s)} = F_s(s) = \frac{K_1G_1(s)}{1 + K_1G_1(s)K_2G_2(s)} = \frac{1}{[1/K_1G_1(s)] + K_2G_2(s)} \tag{4.5}$$

The Laplace transform of the error $e(t)$ defined by Eq. (3.11), if denoted by $E(s)$, is then

$$\frac{E(s)}{Y(s)} = \frac{X(s) - Y(s)}{Y(s)} = \frac{1}{F_s(s)} - 1 = \frac{1}{K_1G_1(s)} - [1 - K_2G_2(s)] \quad (4.6)$$

For *simple feedback servomechanisms*, as shown in Fig. 4.3, the transfer function of the feedback link $F_2(s)$ is simply unity; *i.e.*, the output is only

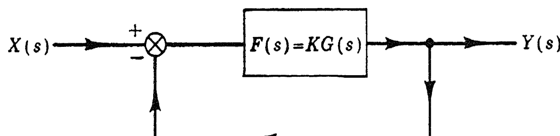


Fig. 4.3

measured, not modified for the feedback control. Then Eqs. (4.6) and (4.7) simplify to

$$\frac{Y(s)}{X(s)} = F_s(s) = \frac{KG(s)}{1 + KG(s)} = \frac{1}{[1/KG(s)] + 1} \quad (4.7)$$

and

$$\frac{E(s)}{Y(s)} = \frac{1}{KG(s)} \quad (4.8)$$

The first requirement of a servomechanism is stability. This means that the output $y(t)$ should be damped, except possibly for steady sinusoidal motion. Our analysis in Sec. 2.4 shows, however, that the condition of stability is mathematically equivalent to the statement that $F_s(s)$ should have no poles in the right-half s plane, where the real part of s is positive. For the general feedback servomechanism, as shown by Eq. (4.6), the poles of $F_s(s)$ are zeros of

$$\frac{1}{F_s(s)} = \frac{1}{K_1G_1(s)} + K_2G_2(s) \quad (4.9)$$

For the simple feedback servomechanism, as shown by Eq. (4.8), the poles of $F_s(s)$ are zeros of

$$\frac{1}{F_s(s)} = \frac{1}{KG(s)} + 1 \quad (4.10)$$

Therefore the first design criterion of feedback servomechanisms is

(*a*) *The function* $1/F_s(s)$, *given by Eqs.* (4.9) *and* (4.10), *should not have zeros in the right-half* s *plane.*

The second requirement of a servomechanism is quick response. If s_r is a pole of $F_s(s)$, then the analysis in Sec. 2.4 shows that the output has the component e^{s_rt}. The quickness of response is thus determined by the magnitude of s_r. The larger the magnitude of s_r, the shorter the time scale and thus the quicker the response. Therefore the second design criterion of a feedback servomechanism is

(*b*) *The zeros of* $1/F_s(s)$, *given by Eqs.* (4.9) *and* (4.10), *should all be of large magnitude and lie sufficiently to the left of the imaginary axis of the s plane.*

If the feedback servomechanism is designed for controlling the output to follow the input signal, the steady-state output after the removal of transients should be made as close as possible to the input. Therefore, for such "positional" controls, there is the third requirement that the ratio $E(0)/Y(0)$ between the steady-state error and the steady-state output should be as small as possible. This condition can be translated into a condition on the gains of the transfer functions by using Eqs. (4.6) and (4.8). Thus

(*c*) *For positional control, accuracy of control requires for general servomechanisms, Eq.* (4.7),

$$\frac{1}{K_1} - [1 - K_2] \sim 0 \tag{4.11}$$

and for simple servomechanisms, Eq. (4.9),

$$K \gg 1 \tag{4.12}$$

The conditions (*a*), (*b*), and (*c*) are the design criteria of feedback servomechanisms. In practice, it is usually difficult to satisfy conditions (*b*) and (*c*) as fully as desired, and a compromise has generally been made. We shall see this in the following sections.

4.3 Method of Nyquist. Since, as stated before, the transfer functions are usually ratios of two polynomials in *s*, criterion (*a*) of the last section is generally equivalent to specifying the nonexistence of roots with positive real parts for a polynomial. This is a classic question and is answered by E. J. Routh using the so-called Routh inequalities, involving the coefficients of the polynomial under investigation. This method, however, is not favored by control engineers, because of the obscure manner of the variation of the Routh inequalities with changes in the coefficients. Engineers prefer a method of analysis which uses the transfer functions written in Eqs. (4.9) and (4.10) directly without further modification; because these transfer functions are the immediate information possessed and are understood "physically" by the engineer.

Such a method was devised by H. Nyquist. The Nyquist method is based upon a theorem due to Cauchy for an analytical function $f(s)$, where *s* is a complex variable:[1]

If $f(s)$ *has n zeros and m poles within a closed path C, then as s travels*

[1] See for instance Whittaker and Watson, "Modern Analysis," Sec. 6.31, p. 119, Cambridge-Macmillan, 1943.

along C once in a clockwise direction, the vector $f(s)$ carries out $n - m$ clockwise revolutions about the origin.

To apply this very powerful theorem to our problem, we have chosen the path C to enclose the whole right-half s plane, where zeros with positive real parts would lie. Such a path is shown in Fig. 4.4 and consists of the imaginary axis and a semicircle to the right with the radius $R \to \infty$. Take first the simpler case, the case of a simple feedback servomechanism. We note from Eq. (4.10) that the poles of $1/F_s(s)$ are zeros of $G(s)$. Let the number of zeros of $G(s)$ in the right-half s plane be m. Then $1/F_s(s)$ has m poles in C. Therefore, in order for $1/F_s(s)$ to have no zeros in the right-half s plane, $1/F_s(s)$ has to carry out m counterclockwise revolutions around the origin when s describes the contour C of Fig. 4.4 with $R \to \infty$. But from Eq. (4.10), it is easily seen that this is equivalent to requiring $1/KG(s)$ to carry out m counterclockwise revolutions around the point -1. But since K is a constant, the above criterion is the same as requiring $1/G(s)$ to carry out m counterclockwise revolutions around the point $-K$. Needless to say, when $G(s)$ has no zero in the right-half s plane, or $m = 0$, then the Nyquist stability criterion requires the vector $1/G(s)$ to make no revolution around the point $-K$.

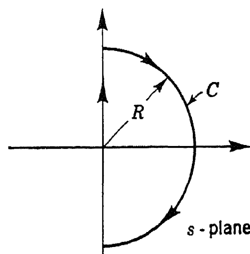


FIG. 4.4

Let us illustrate the application of the method by taking the simple transfer function

$$\left.\begin{aligned} G(s) &= \frac{1}{s(1 + \tau_1 s)(1 + \tau_2 s)} \\ \text{Then} \qquad 1/G(s) &= s(1 + \tau_1 s)(1 + \tau_2 s) \end{aligned}\right\} \tag{4.13}$$

First consider the part of the path C of Fig. 4.4 along the imaginary axis, where

$$\frac{1}{G(i\omega)} = i\omega(1 + i\tau_1\omega)(1 + i\tau_2\omega)$$

At $\omega = 0$, $1/G(i\omega) = i0$. As $\omega \to +\infty$, $1/G(i\omega) \to -i\infty$. Therefore as ω increases from 0 to ∞, the vector $1/G(i\omega)$ increases in magnitude, and its phase angle increases from $\pi/2$ to $3\pi/2$. For negative ω, the curve traced by the end point of the vector $1/G(i\omega)$ is simply the reflection of the curve for positive ω about the real axis, as required by Eq. (3.17). Thus as s traces the imaginary axis, $1/G(i\omega)$ traces the curve *aboc* shown in Fig. 4.5.

As s traces the large semicircle shown in Fig. 4.4, $1/G(s) \sim s^3$. Then as s rotates from $i\infty$ to $-i\infty$ clockwise, $1/G(s)$ will also rotate clockwise, but three times as fast. This part of $1/G(s)$ is thus represented by the

curve c to a in Fig. 4.5. From this figure, it is seen that if $K = K_I$ as indicated in the figure, then the vector $1/G(s)$ will make no net revolutions around the point $-K_I$. Since the function $G(s)$, given by Eq. (4.13), has no zero, this means that the feedback system will be stable. If $K = K_{II}$ as indicated, then the vector $1/G(s)$ will make two net clockwise revolutions around the point $-K_{II}$. Therefore, with this larger value of K, the feedback servomechanism will be unstable. In fact, there will be two poles of $F_s(s)$ with positive real parts. The transition point from stability to instability is the point b. Stable values of the gain K must lie between the origin and this point.

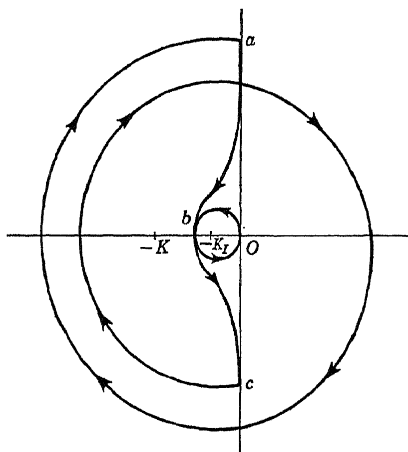


Fig. 4.5

For a general feedback servomechanism, the question is whether there is any zero of the expression $1/F_s(s)$ given by Eq. (4.9) in the right-half s plane. To use this expression directly for the Cauchy theorem is inconvenient, because then we have to add two vectors $1/K_1G_1(s)$ and $K_2G_2(s)$. Now let $G_1(s)$ and $G_2(s)$ have m_1 and m_2 zeros, respectively, in the right-half s plane. The respective numbers of poles in the right-half s plane are n_1 and n_2. Then it is evident that the number of poles of $1/F_s(s)$ in the right-half s plane is $m_1 + n_2$. Now let us divide $1/F_s(s)$ by $K_2G_2(s)$. This operation will introduce m_2 poles and n_2 zeros into the expression. But there is a possibility that some of the zeros may be the same as the poles, and thus both are removed. Let the number of zeros and poles thus removed be α. Now

$$\frac{1}{F_s(s)K_2G_2(s)} = \frac{1}{K_1K_2G_1(s)G_2(s)} + 1 \tag{4.14}$$

The number of poles of $1/F_s(s)K_2G_2(s)$ in the right-half s plane is the number of poles of $1/K_1K_2G_1(s)G_2(s)$ and is thus equal to $m_1 + m_2$. Now

let us assume that there is no zero of $1/F_s(s)$ in the right-half s plane, *i.e.*, the feedback system is stable. Then the numbers of zeros and poles of various expressions are as listed in Table 4.1. It can then be easily deduced that $n_2 - \alpha = 0$, and $1/F_s(s)K_2G_2(s)$ also has no zeros in the

TABLE 4.1

Expression	Number in right-half s plane of	
	Zeros	Poles
$G_1(s)$	m_1	n_1
$G_2(s)$	m_2	n_2
$1/F_s(s)$	0	$m_1 + n_2$
$1/F_s(s)K_2G_2(s)$	$n_2 - \alpha$	$m_1 + n_2 + m_2 - \alpha = m_1 + m_2$

right-half s plane. Hence as s traces the path C specified in Fig. 4.4, the vector $1/F_s(s)K_2G_2(s)$ should make $-(m_1 + m_2)$ clockwise revolutions around the origin. By referring to Eq. (4.14), it is seen that this condition of stability is equivalent to requiring the vector $1/K_1K_2G_1(s)G_2(s)$ to make $m_1 + m_2$ counterclockwise revolutions around the point -1. Or we may require the vector $1/G_1(s)G_2(s)$ to make $m_1 + m_2$ counterclockwise revolutions around the point $-(K_1K_2)$. This is the Nyquist criterion for stability of a general feedback servomechanism.

The essential part of the locus of the path of integration in the Nyquist method is the part where $s = i\omega$, as is clearly demonstrated by our example in Fig. 4.5. Therefore the stability problem can be solved by using directly the data on the frequency response of the forward link and the feedback link. Since the frequency response of the elements in the system is often determined experimentally, a method allowing the direct application of experimental information has advantages. This is the merit of the Nyquist method. Its drawback is the uncertainty about the degree of stability. That is, if stable, what is the magnitude of damping? To answer this question, we may modify the criterion to require no zeros of $1/F_s(s)$ to the right of a line parallel to the imaginary axis in the s plane but displaced to the left. The distance $-\lambda$ between this line and the imaginary axis specifies the minimum amount of damping. The Nyquist criterion can again be used, with the proper modifications on the path C, for s and the numbers of zeros m, m_1, and m_2. However, to carry out this test, we have to know the value of the transfer functions, not at $s = i\omega$, but at $s = -\lambda + i\omega$. Hence information on the frequency response can no longer be used directly. Then the method of Nyquist loses its main advantage. In fact, a different approach to the question devised by W. R. Evans[1] is much better. We shall discuss this method in the following section.

4.4 Method of Evans. Let us consider first the case of simple feedback servomechanisms. Then the basic question is to find the roots of the equation

$$0 = \frac{1}{F_s(s)} = 1 + \frac{1}{KG(s)} \tag{4.15}$$

with $G(s)$ given. The Evans method determines such roots as functions of the gain K and is thus called the *root-locus method.* When this is done, any set of specifications on the roots gives a proper choice of the magnitude of K. This method thus goes much beyond the mere satisfaction of criterion (*a*) of Sec. 4.2 and actually solves the design problem for all three criteria stated in that section.

Now let $G(s)$ be specified by its zeros $p_1, p_2, \ldots, p_m$ and its poles $q_1, q_2, \ldots, q_n$. Then from the definition of gain given by Eqs. (3.16), (3.21), and (3.23),

$$G(s) = A\,\frac{(s - p_1)(s - p_2)\cdots(s - p_m)}{(s - q_1)(s - q_2)\cdots(s - q_n)} \tag{4.16}$$

where

$$A = \frac{(-q_1)(-q_2)\cdots(-q_n)}{(-p_1)(-p_2)\cdots(-p_m)}$$

For physical systems, the polynomials in the numerator and the denominator of $G(s)$ have real coefficients. Then the p's are either real or form complex conjugate pairs. Similarly, the q's are either real or form complex conjugate pairs. Therefore A is always real. For engineering systems, usually things are so arranged as to make A not only real but also positive. Hereafter, then, we shall consider A to be real and positive. Generally, the denominator of $G(s)$ is of equal or higher order than the numerator, that is, $n \geq m$. Let us express each of the factors in Eq. (4.16) in vector form:

$$\left.\begin{aligned} s - p_1 &= P_1 e^{i\varphi_1} \\ s - p_2 &= P_2 e^{i\varphi_2} \\ &\cdots\cdots\cdots \\ s - p_m &= P_m e^{i\varphi_m} \end{aligned}\right\} \tag{4.17}$$

$$\left.\begin{aligned} s - q_1 &= Q_1 e^{i\theta_1} \\ s - q_2 &= Q_2 e^{i\theta_2} \\ &\cdots\cdots\cdots \\ s - q_n &= Q_n e^{i\theta_n} \end{aligned}\right\} \tag{4.18}$$

The vector $P_r e^{i\varphi_r}$ goes from p_r to s. The vector $Q_r e^{i\theta_r}$ goes from q_r to s. s is the variable point in the complex s plane. By using Eqs. (4.17) and (4.18), $G(s)$ can be written as

$$G(s) = A\,\frac{(P_1 e^{i\varphi_1})(P_2 e^{i\varphi_2})\cdots(P_m e^{i\varphi_m})}{(Q_1 e^{i\theta_1})(Q_2 e^{i\theta_2})\cdots(Q_n e^{i\theta_n})} \tag{4.19}$$

Since A is real and positive, we can write Eq. (4.19) as

$$G(s) = Re^{i\Theta} \tag{4.20}$$

where

$$R = A\frac{(P_1P_2 \cdots P_m)}{(Q_1Q_2 \cdots Q_n)} \tag{4.21}$$

and

$$\Theta = (\varphi_1 + \varphi_2 + \cdots + \varphi_m) - (\theta_1 + \theta_2 + \cdots + \theta_n) \tag{4.22}$$

Since the P's and Q's are magnitudes of vectors defined by Eqs. (4.17) and (4.18), they are positive. Therefore R is positive. The basic equation for the roots of the inverse system transfer function, Eq. (4.15), is thus

$$\frac{e^{-i\Theta}}{KR} = -1$$

Therefore, to satisfy this equation, we must have

$$KR = 1 \tag{4.23}$$

and

$$\Theta = \pm\pi \tag{4.24}$$

The Evans method consists of two steps: The first step is to determine all s's that satisfy the appropriate angle condition of Eq. (4.24). Then, knowing such a root locus, we can compute R and hence K, by Eq. (4.23), for each point on the root locus. Evans has developed a number of useful rules for plotting the root locus. We shall now explain these rules.

Rule 1. For $K = 0$, Eq. (4.15) shows that $G(s) \to \infty$. Thus for $K = 0$, the roots of $1/F_s(s)$ are poles of $G(s)$, or the root locus starts at the poles of $G(s)$. These poles of $G(s)$ will be denoted by a dot in the s plane.

Rule 2. For $K \to \infty$, $G(s) \to 0$. Thus for $K \to \infty$, the root locus could be the zeros of $G(s)$. We shall denote the zeros of $G(s)$ by a small circle in the s plane. But if $n > m$, the number of zeros of $G(s)$ is less than the number of zeros of $1/F_s(s)$. However, in that case, $G(s) \to 0$ as $s \to \infty$. Therefore the missing roots are supplied by $s = \infty$. Furthermore, for very large s,

$$G(s) \sim \frac{A}{s^{n-m}}$$

Therefore Eq. (4.15) can be approximated by

$$s^{n-m} \approx -KA$$

Thus the asymptotes of the root locus have the phase angles

$$\frac{\pi}{n-m} + \frac{2k\pi}{n-m} \qquad k = 1, 2, 3, \ldots \tag{4.25}$$

Rule 3. The root locus along the real axis is along alternate segments connecting zeros and poles of $G(s)$ located on the real axis, starting with the one farthest to the right.

This rule may be easily verified by considering any point s on the real axis. The angles to this point from a pair of complex conjugate zeros or poles are $+\varphi$ and $-\varphi$ or $+\theta$ and $-\theta$, respectively. Thus their sum is zero. The angle to this point from a pole or zero on the axis is 0 for all poles or zeros to the left of s, and it is π for all poles or zeros to the right of s. Thus, the sum is π if there is an odd number of poles and zeros of $G(s)$ to the right of s.

Rule 4. If there is a breakaway of the root locus from the real axis, the point of breakaway may be estimated from the condition that, for a small displacement $\Delta\omega$ from the axis, the increase in angle due to the poles and zeros of $G(s)$ on the axis to the left must be just balanced by the effect of those to the right.

Example: Consider the transfer function

$$G(s) = \frac{(0.001)(2)(6)}{(s + 0.001)(s + 2)(s + 6)} \tag{4.26}$$

At $K = 0$, the locus starts from -0.001, -2, and -6 on the real axis. Sections of the locus lie between -0.001 and -2, and between -6 and $-\infty$. Here $m = 0$, $n = 3$. Therefore the phase angles for the asymptotes, according to Eq. (4.25), are $+\pi/3$, $-\pi/3$, and π. Breakaway from the real axis occurs at λ_1, between -0.001 and -2. By applying Rule 4, we have

$$\frac{\Delta\omega}{\lambda_1 + 0.001} + \frac{\Delta\omega}{\lambda_1 + 2} + \frac{\Delta\omega}{\lambda_1 + 6} = 0$$

or

$$(\lambda_1 + 2)(\lambda_1 + 6) + (\lambda_1 + 0.001)(\lambda_1 + 6) + (\lambda_1 + 0.001)(\lambda_1 + 2) = 0$$

Hence

$$3\lambda_1^2 + 16.002\lambda_1 + 12.008 = 0$$

Therefore

$$\lambda_1 = -\frac{16.002}{6} - \sqrt{\left(\frac{16.002}{6}\right)^2 - \frac{12.008}{3}} = -0.904$$

Rule 5. The point at which the root locus crosses the imaginary axis into the right-half s plane can often be estimated by taking advantage of the properties of the right angle.

Example: Let us consider the same transfer function, of Eq. (4.26). Away from the origin, it can be very closely approximated by

$$G(s) \approx \frac{(0.001)(2)(6)}{s(s + 2)(s + 6)}$$

Then, as shown in Fig. 4.6, $\theta_1 \approx \pi/2$. Therefore Eq. (4.24) gives

$$\theta = -\pi = -\theta_2 - \theta_3 - \frac{\pi}{2}$$

or

$$\theta_2 + \theta_3 \approx \frac{\pi}{2}$$

But, by referring to the figure,

$$\frac{\pi}{4} = \theta_3 + \beta \qquad \text{and} \qquad \frac{\pi}{4} + \alpha = \theta_2$$

hence

$$\alpha \approx \beta$$

This is the geometrical condition for determining the crossover point U.

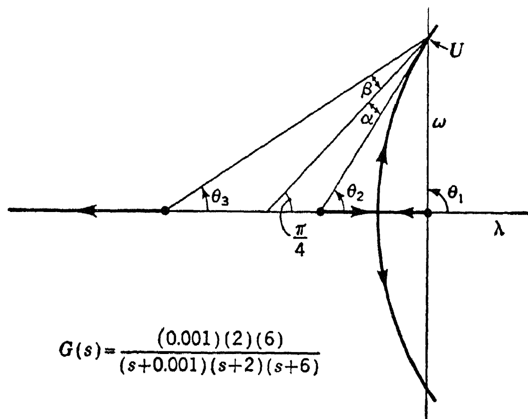


Fig. 4.6

Rule 6. The direction of locus departure from a pole (or locus approach to a zero) may be easily estimated by computing the angle at the pole (or zero) under consideration from all the other poles and zeros in the field.

Example: Figure 4.7 shows the root locus for a transfer function $G(s)$ having two zeros and two poles on the real axis, and a pair of complex conjugate poles. For small displacements away from the pole q_4, the angles φ_1, φ_2, θ_1, θ_2, and θ_3 from other zeros and poles remain constant. Thus the angle θ_4 is given, according to Eq. (4.24), by

$$(\varphi_1 + \varphi_2) - [(\theta_1 + \theta_2 + \theta_3) + \theta_4] = \pi$$

This equation determines θ_4.

These rules give the essential characteristics of the root locus. For intermediate locations, the root locus is found by taking a number of

trial points. The gain K can then be calculated along the path of the root locus. When the desired location of the roots of $1/F_s(s)$ is finally chosen, the corresponding value of K can thus be fixed. The design of the feedback system is then completed.

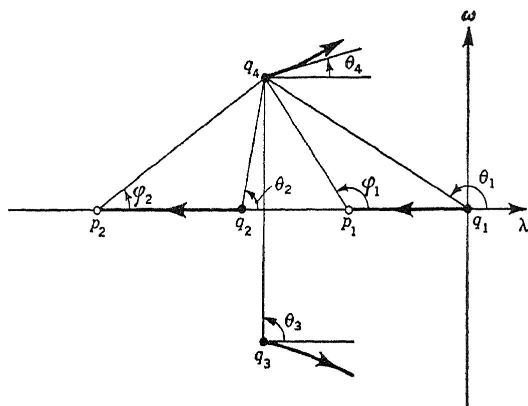


FIG. 4.7

4.5 Hydrodynamic Analogy of Root Locus. By combining Eqs. (4.15) and (4.16), we have

$$\frac{(s - q_1)(s - q_2) \cdots (s - q_n)}{(s - p_1)(s - p_2) \cdots (s - p_m)} = -KA$$

If we take the logarithm of the above equation and then divide the resultant equation by 2π, we have

$$W(s) = \frac{1}{2\pi} \sum_{i=1}^{n} \log (s - q_i) - \frac{1}{2\pi} \sum_{j=1}^{m} \log (s - p_j) = \frac{1}{2\pi} \log KA + i \left(\frac{1}{2}\right) \tag{4.27}$$

Equation (4.27) has many possible physical interpretations. A very illuminating one is to consider $W(s)$ as the complex potential function of a two-dimensional irrotational flow of a perfectly incompressible fluid.[1] If $\phi(\lambda,\omega)$ and $\psi(\lambda,\omega)$ are the potential function and the stream function, respectively, then

$$W(s) = \phi(\lambda,\omega) + i\psi(\lambda,\omega) \tag{4.28}$$

with $s = \lambda + i\omega$. Therefore Eq. (4.27) for the root locus of $1/F_s(s)$ can be interpreted as lines on which the stream function ψ assumes the constant value $\frac{1}{2}$. In the terminology of fluid mechanics, the root locus

[1] See for instance V. L. Streeter, "Fluid Dynamics," McGraw-Hill Book Company, Inc., New York, 1948.

is thus composed of branches of the $\frac{1}{2}$ streamline. The potential function along the streamline changes from point to point and is equal to

$$\frac{1}{2\pi} \log KA$$

Equation (4.27) also shows that the flow is composed of n *sources* of unit strength located at the points $q_1, q_2, \ldots, q_n$ and of m *sinks*, also of unit strength, located at the points $p_1, p_2, \ldots, p_m$. In our graphical representation, sources are indicated by a dot, and sinks by a small circle. With this interpretation, the pattern of root loci in Figs. 4.6 and 4.7 can be immediately "understood."

The hydrodynamic analogy is also very useful in suggesting modifications of the system to achieve a better feedback performance. For instance, a system characterized by the transfer function

$$G(s) = \frac{q_1 q_2}{s(s - q_1)(s - q_2)} \qquad |q_1| < |q_2|$$

may have the disadvantage of being unstable in closed-cycle performance at too low values of the gain K, and thus not being able to satisfy criterion (*c*) of Sec. 4.2. The root locus is similar to that shown in Fig. 4.6. The hydrodynamic analogy immediately suggests that the crossover point U can be moved up by pulling that part of the streamline near U to the left, with a sink p_c close to q_1, and a source q_c near to q_2. Thus the modified transfer function is

$$G(s) = \frac{q_c}{p_c} \frac{(s - p_c)}{(s - q_c)} \frac{q_1 q_2}{s(s - q_1)(s - q_2)}$$

The corresponding root locus is shown in Fig. 4.8. Since $|p_c| < |q_c|$, the additional transfer function put in series with the original transfer function must be that of a lead network, as shown in Sec. 3.3 by Eq. (3.27).

The hydrodynamic analogy also permits us to understand the possibility of speeding up the response of a slow mechanism by using the feedback link. According to criterion (*b*) of Sec. 4.2, fast response requires roots with large magnitudes. Now, for simplicity, suppose we have a linear mechanical system of first order characterized by a small q_1 on the negative part of the real axis. If we put this system in series with a fast-damped electric network characterized by a large q_2 on the negative real axis, the response will not be improved, because we still have the small q_1 root. But if we have closed the feedback cycle, then the streamline pattern, or the root locus, indicates that the smaller root q_1 will increase with increasing gain K toward the larger root q_2. Therefore, with proper choice of the gain K, we can make the roots much larger than q_1 and thus greatly increase the rapidity of response of the system.

The technique of plotting the root locus can also be applied to general feedback servomechanisms. There the problem is to plot the root locus of $1/F_s(s)$ given by Eq. (4.19). Thus the condition for the root locus is

$$\frac{1}{K_1G_1(s)} = -K_2G_2(s)$$

Since the roots of $1/F_s(s)$ are different from the zeros of $G_2(s)$, we can

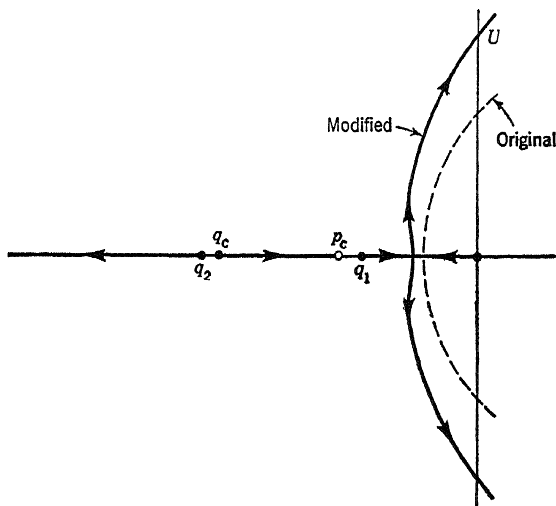


FIG. 4.8

divide the above equation by $G_2(s)/K_1$, and thus

$$\frac{1}{G_1(s)G_2(s)} = -K_1K_2 \tag{4.29}$$

Therefore, if we put

$$\left.\begin{aligned} G(s) &= G_1(s)G_2(s) \\ K &= K_1K_2 \end{aligned}\right\} \tag{4.30}$$

and then compare Eq. (4.29) with Eq. (4.15), we see that the problem of finding the root locus of the general feedback servomechanism is reduced to that of the simple feedback servomechanism discussed previously. In fact, our careful analysis of the application of the Nyquist method to general feedback servomechanisms in Sec. 4.3 shows that the method of reduction given by Eq. (4.30) can also be used there. Therefore, as far as finding the qualitative performance specified by criteria (*a*), (*b*), and (*c*) of Sec. 4.2 is concerned, there is no difference between the simple feedback servomechanism and the general feedback servomechanism, if the relations of Eq. (4.30) are borne in mind. Only when the quantitative performance of the system is required must the differences

in the system transfer functions $F_s(s)$ as shown by Eqs. (4.3) and (4.7) be properly recognized.

4.6 Method of Bode. At the point U where the root locus crosses over to the right-half s plane, the root is by definition purely imaginary, say $i\omega^*$. In other words, Eq. (4.15) is satisfied by $s = i\omega^*$, or

$$KG(i\omega^*) = F(i\omega^*) = -1 = 1 \cdot e^{-i\pi}$$

Therefore, the critical condition of transition from stability to instability occurs if the amplitude M of the frequency response is equal to unity and, simultaneously, the phase angle θ of the frequency response is equal to $-\pi$. This critical condition can also be deduced from the Nyquist criterion, which specifies the critical point as -1 in the $1/F(i\omega)$ diagram. In fact, by studying a typical example, such as Fig. 4.5, it will be seen that for stability the $1/F(i\omega)$ curve must encircle the -1 point. Since the magnitude of $1/F(i\omega)$ generally increases with increasing ω, encirclement can be ensured by requiring that the magnitude of $1/F(i\omega)$ be larger than 1 when the phase angle of $1/F(i\omega)$ is equal to π. This is equivalent to saying that M should be less than one when θ is equal to $-\pi$. Or, we say that θ should be larger than $-\pi$ when $M = 1$. This condition for stability is the basis of the method of Bode: the frequency at which the amplitude M of the frequency response is equal to 1 is called the point of *gain crossover*. The difference of θ and $-\pi$ is called the *phase margin*. The Bode criterion for stability is thus stated as a phase margin of 30 to 50 degrees at gain crossover. In a Bode diagram, the point of gain crossover is the frequency for $\log_{10} M = 0$, and the Bode criterion can be easily tested.

The Bode method is similar to the Nyquist method in that the information on the frequency response can be used directly. This advantage of simplicity is counterbalanced by the disadvantage, in comparison with the Evans root-locus method, of not being able to know the *degree* of stability. R. M. Osborn[1] tried to remedy this situation by giving a semiempirical rule to calculate the damping coefficient ζ for the most critical root. His formula is

$$\zeta \approx \frac{1}{60} \frac{\alpha}{m} \tag{4.31}$$

where α is the phase margin in degrees at gain crossover, and m is the slope of $\log_{10} M$ against ω at the gain crossover. The unit of time for ζ is the same as that for ω. Thus, if $\alpha = 30$ degrees and $m = 1.7$, then $\zeta \approx 1/(2 \times 1.7) = 0.3$.

4.7 Designing the Transfer Function. The various methods discussed in the previous sections for determining the stability of feedback

[1] R. M. Osborn, paper presented at Summer Meeting, IRE, San Francisco, August. 1949.

servomechanisms are mainly methods of analysis. They are partly methods of synthesis, *i.e.*, methods of designing the transfer function, but only *in so far as* they fix the range of possible values of the gain K. Of course, both methods can suggest changes in the structure of the transfer function to improve performance. This is particularly so in the case of the root-locus method. However, how to realize these desired changes in the transfer function by modifying the physical elements of the system is mainly an art in servomechanism engineering practice.

Only in one aspect of this synthesis problem is a general solution known. This is the problem of designing an electric circuit composed of

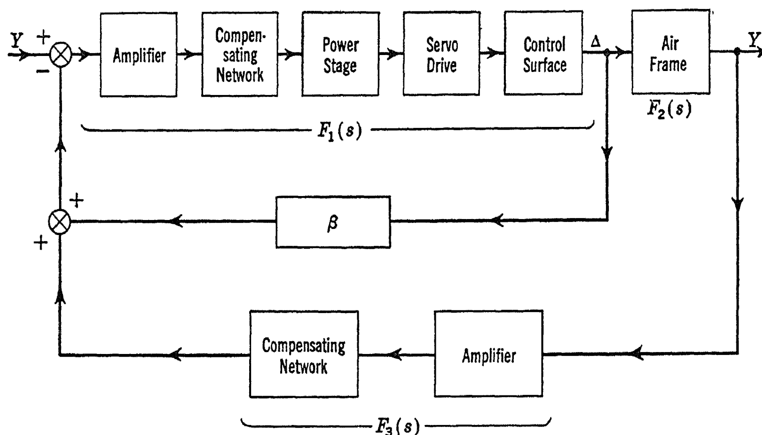


Fig. 4.9

resistances and capacitances, an *RC* circuit, such that the transfer function of this electric circuit has the specified zeros and poles. Since such a circuit has great flexibility and is used very often to "compensate" the transfer-function characteristics of other elements in the system, and since the desired modifications of the transfer function can indeed often be put in terms of additional zeros and poles, the general solution of such a problem is very important. Important contributions to this problem were made by E. A. Guillemin[1] and L. Weinberg.[2] We shall not pursue the subject here, but only emphasize the possibility of synthesizing an *RC* circuit of very complex specified properties.

4.8 Multiple-loop Servomechanisms. The servomechanisms discussed thus far are single-loop servomechanisms. Engineering practice often calls for much more complicated systems. For instance, Fig. 4.9 is the block diagram of a typical control system[3] for an airplane rotating

[1] E. A. Guillemin, *J. Math. and Phys.*, **28**, 22–44 (1949).
[2] L. Weinberg, *J. Appl. Phys.*, **24**, 207–216 (1953).
[3] L. Becker, *Aeronaut. Eng. Rev.*, September, 1951, p. 17.

about a single axis. The inner loop is the so-called control surface-position feedback or "follow-up." If the inner loop is not closed, we have the usual feedback control, and

$$\frac{Y(s)}{X(s)} = \frac{F_1(s)F_2(s)}{1 + F_1(s)F_2(s)F_3(s)} \tag{4.32}$$

If both loops are closed, then

$$\Delta(s) = F_1(s)[X(s) - \beta\Delta(s) - F_3(s)Y(s)]$$

and

$$Y(s) = F_2(s)\Delta(s)$$

Therefore

$$\frac{\Delta(s)}{X(s)} = \frac{F_1(s)}{1 + \beta F_1(s) + F_1(s)F_2(s)F_3(s)} \tag{4.33}$$

and

$$\frac{Y(s)}{X(s)} = \frac{F_1(s)F_2(s)}{1 + \beta F_1(s) + F_1(s)F_2(s)F_3(s)} \tag{4.34}$$

The stability and response of the control system then depend upon the zeros of the expression $1 + \beta F_1(s) + F_1(s)F_2(s)F_3(s)$.

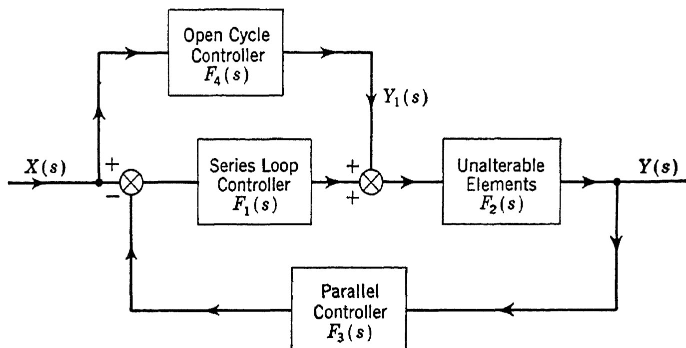


FIG. 4.10

One of the difficulties of designing a good control system is to have accurate control and hence large gain K together with fast response and satisfactory damping. This led to the idea of combining the closed-cycle control with open-cycle control, proposed by J. R. Moore.[1] Consider the system represented by Fig. 4.10, where the closed-cycle control and the open-cycle control are put in parallel. We have thus

$$\left.\begin{aligned} F_4(s)X(s) &= Y_1(s) \\ \text{and} \qquad Y(s) &= F_2(s)\{Y_1(s) + F_1(s)[X(s) - F_3(s)Y(s)]\} \end{aligned}\right\} \tag{4.35}$$

[1] J. R. Moore, *Proc. IRE*, **39**, 1421–1432 (1951).

Solving for the output $Y(s)$, we have

$$\frac{Y(s)}{X(s)} = \frac{F_2(s)F_4(s) + F_1(s)F_2(s)}{1 + F_1(s)F_2(s)F_3(s)} \tag{4.36}$$

The stability and the speed of response of the system are thus established by the zeros of $1 + F_1(s)F_2(s)F_3(s)$. Since $F_2(s)$ is fixed, the design problem is to find the proper transfer functions $F_1(s)$ and $F_3(s)$. The actual response, in particular the steady-state error, is dependent upon the additional transfer function $F_4(s)$ of the open-cycle controller. Therefore the feedback loop is designed primarily for stability and dynamic response, while the steady-state or "synchronizing" operations are largely taken care of by the open-cycle portion of the system.

When there are many variables to be controlled simultaneously and when these controlled variables are also coupled, as in a steam power plant, then the system diagram has many loops with a complicated feedback scheme.[1] An extreme example of such complex systems is, perhaps, the automatic control and guidance system for airplanes.[2] The analysis of such a system, although following the same principles as explained in this chapter for simple servomechanisms, can hardly be done without recourse to analog computers. But this is only engineering development work: the process of going from principles to practice.

[1] See for instance J. Hänny, "Regelung Theorie," A. G. Gebr. Leemann Co., Zürich, 1946.

[2] J. B. Rea, *Aeronaut. Eng. Rev.*, November, 1951, p. 39.

CHAPTER 5

NONINTERACTING CONTROLS

For complex systems with several controlled quantities and with interaction between these controlled quantities, a new design criterion generally has to be introduced. This is the criterion of noninteraction. For instance, the variables of a turbojet engine with afterburning are the engine speed, the fuel injection rate to the combustion chambers, the fuel injection rate to the afterburner, and the cross-sectional area of the tail-pipe opening. However, the operation of this engine may be based upon specific settings of the speed, the fuel rate to the combustion chambers, and the fuel rate to the afterburner. If this is the case, it is obvious that one of the design criteria for the servocontrol of the system is the independence of the three different control settings: a change in fuel rate to the afterburner should not change the engine speed, and a change in engine speed should not require a change in fuel rate to the combustion chambers. The key to this particular design problem is then the proper manipulation of the tail-pipe opening with respect to the other variables and the proper design of the control servos. The purpose of this chapter is to give a general method for designing such noninteracting controls for systems of arbitrary complexity. This general method was first given by A. S. Boksenbom and R. Hood.[1]

5.1 Control of a Single-variable System. Let us consider first a simple system with one controlled output $y(t)$ and one control setting, or input, $x(t)$. Their Laplace transforms are $Y(s)$ and $X(s)$. Consider the control designed according to Fig. 5.1. $E(s)$ is the "engine" transfer function, $L(s)$ is the instrument transfer function, $S(s)$ is the servo transfer function, and $C(s)$ is the "control" transfer function. Only $C(s)$ can be changed easily by the designer. The system is slightly different from the simple servomechanism of Fig. 4.2, in that an arbitrary disturbance $V(s)$ is introduced between the servo and the engine to account for accidental outside influences.

The relation between input $W(s)$ to the engine and the output $Y(s)$ is

$$Y(s) = E(s)W(s) = E(s)[S(s)U(s) + V(s)] \tag{5.1}$$

$U(s)$ is the output of the control transfer function and is in turn given by

$$U(s) = C(s)[X(s) - Z(s)] = C(s)[X(s) - L(s)Y(s)] \tag{5.2}$$

[1] A. S. Boksenbom and R. Hood, *NACA TR* 980 (1950).

By eliminating $U(s)$ from Eqs. (5.1) and (5.2), we have

$$Y(s) = \frac{E(s)S(s)C(s)}{E(s)S(s)C(s)L(s) + 1} X(s) + \frac{E(s)}{E(s)S(s)C(s)L(s) + 1} V(s) \quad (5.3)$$

This is the equation for the Laplace transform of the output under appropriate initial conditions for $y(t)$ and $x(t)$. Except for the second term, involving the disturbance $V(s)$, Eq. (5.3) is the same as the previous relation, Eq. (4.3), for simple servomechanisms. The analysis of the performance of the system can also be carried out in a similar way. However, for more complicated systems, this scheme has to be generalized. We shall do this presently.

FIG. 5.1

5.2 Control of a Many-variable System. Let the number of outputs $Y_1(s), Y_2(s), \ldots, Y_\nu(s), \ldots, Y_i(s)$ of the engine be i and the number of inputs $W_1(s), W_2(s), \ldots, W_k(s), \ldots, W_n(s)$ be n. Then the generalization of Eq. (5.1) is

$$\left.\begin{aligned} Y_1(s) &= E_{11}(s)W_1(s) + E_{12}(s)W_2(s) + \cdots + E_{1n}(s)W_n(s) \\ Y_2(s) &= E_{21}(s)W_1(s) + E_{22}(s)W_2(s) + \cdots + E_{2n}(s)W_n(s) \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Y_i(s) &= E_{i1}(s)W_1(s) + E_{i2}(s)W_2(s) + \cdots + E_{in}(s)W_n(s) \end{aligned}\right\} \quad (5.4)$$

Each $E_{jk}(s)$ is the transfer function which, when operated on the input $W_k(s)$, gives a component of the output $Y_j(s)$. $E_{jk}(s)$ is then generally a ratio of two polynomials of s, either obtained theoretically from analyzing the engine characteristics, or determined experimentally through the frequency response. Equation (5.4) can be compressed into

$$Y_\nu(s) = \sum_{k=1}^{n} E_{\nu k}(s)W_k(s) \quad (5.5)$$

The array of the quantities $E_{\nu k}(s)$ can be conveniently called the engine transfer-function matrix E. We may then consider that the inputs

$W_k(s)$ "enter" the matrix as columns and that the outputs $Y_\nu(s)$ "leave" the matrix as rows, as indicated in Fig. 5.2. We shall be concerned with the cases where the number of inputs is greater than the number of outputs, that is, $n \geq i$. Therefore the matrix E is rectangular with more columns than rows. For later use, a square matrix obtained by using only the first i columns is denoted by E^*.

Since the number of inputs to the engine is greater than the number of outputs, the system behavior is not determined by merely giving the settings $X_j(s)$, where $j = 1, \ldots, i$, for the outputs $Y_\nu(s)$, but in addition the settings $\Xi_\mu(s)$ for the variables $W_\mu(s)$, where $\mu = i + 1, \ldots, n$, must also be specified. The controlled quantities are then the outputs $Y_\nu(s)$, for $\nu = 1, \ldots, i$, and the $n - i$ engine inputs $W_\mu(s)$. If the

$$
\begin{array}{cccc|cc}
W_1 & W_2 & & W_i & & W_n \\
E_{11} & E_{12} & \cdots & E_{1i} & \cdots & E_{1n} \rightarrow Y_1 \\
E_{21} & E_{22} & \cdots & E_{2i} & \cdots & E_{2n} \rightarrow Y_2 \\
E_{j1} & E_{j2} & \cdots & E_{ji} & \cdots & E_{jn} \rightarrow Y_j \\
E_{i1} & E_{i2} & \cdots & E_{ii} & \cdots & E_{in} \rightarrow Y_i \\
\multicolumn{4}{c}{E^*} & &
\end{array}
$$

Fig. 5.2

measured values of $W_\mu(s)$ after the instrument are denoted by $\Upsilon_\mu(s)$, then the errors are $\Xi_\mu(s) - \Upsilon_\mu(s)$. The errors of the engine output are defined as the differences $X_\nu(s) - Z_\nu(s)$, where $Z_\nu(s)$ for $\nu = 1, \ldots, i$, are the measured values of the output after the instrument, as shown in Fig. 5.1. The function of the control is to take these errors as inputs and to generate correction signals $U_k(s)$ for the servos. This is the feedback link. In the present generalized control system, the correction signals $U_k(s)$ are made to depend linearly upon all errors. Since there are n error signals, there are n correction signals; k thus ranges from 1 to n. Thus

$$
\left.
\begin{aligned}
U_1(s) &= C_{11}(X_1 - Z_1) + C_{12}(X_2 - Z_2) + \cdots + C_{1i}(X_i - Z_i) \\
&\qquad + C'_{1,i+1}(\Xi_{i+1} - \Upsilon_{i+1}) + \cdots + C'_{1n}(\Xi_n - \Upsilon_n) \\
U_2(s) &= C_{21}(X_1 - Z_1) + C_{22}(X_2 - Z_2) + \cdots + C_{2i}(X_i - Z_i) \\
&\qquad + C'_{1,i+1}(\Xi_{i+1} - \Upsilon_{i+1}) + \cdots + C'_{2n}(\Xi_n - \Upsilon_n) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
U_n(s) &= C_{n1}(X_1 - Z_1) + C_{n2}(X_2 - Z_2) + \cdots + C_{ni}(X_i - Z_i) \\
&\qquad + C'_{n,i+1}(\Xi_{i+1} - \Upsilon_{i+1}) + \cdots + C'_{nn}(\Xi_n - \Upsilon_n)
\end{aligned}
\right\} \qquad (5.6)
$$

where we have separated the control matrix into C and C' to indicate that two kinds of error signals are involved. Equation (5.6) can be compressed into

$$U_k(s) = \sum_{\nu=1}^{i} C_{k\nu}(X_\nu - Z_\nu) + \sum_{\mu=i+1}^{n} C'_{k\mu}(\Xi_\mu - \Upsilon_\mu) \quad \begin{matrix} \nu = 1, \ldots, i \\ k = 1, \ldots, n \end{matrix} \tag{5.7}$$

Each of $C_{k\nu}$ and $C'_{k\mu}$ is, of course, a ratio of two polynomials in s. Equations (5.6) and (5.7) can also be represented graphically as shown in Fig. 5.3.

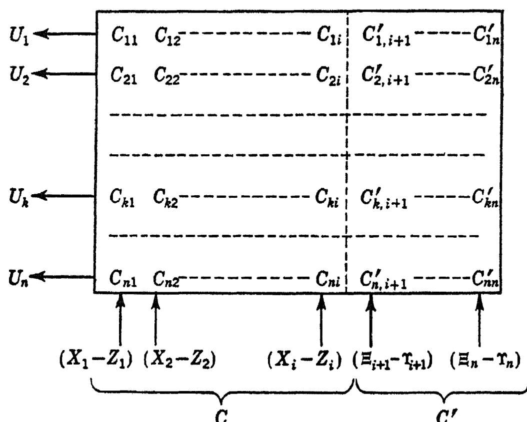


Fig. 5.3

The measured values of $Z_\nu(s)$ and $\Upsilon_\mu(s)$ are related to $Y_\nu(s)$ and $W_\mu(s)$ by the transfer functions $L_{\nu\nu}(s)$ and $L_{\mu\mu}(s)$ of the instruments:

$$Z_\nu(s) = L_{\nu\nu}(s) Y_\nu(s) \tag{5.8}$$

$$\Upsilon_\mu(s) = L_{\mu\mu}(s) W_\mu(s) \tag{5.9}$$

The correction signals will act individually on the servos, and the outputs of the servo when combined with the accidental outside disturbances $V_k(s)$ give the inputs $W_k(s)$ to the engine. If $S_{kk}(s)$ are the transfer functions of the servos, then

$$W_k(s) = S_{kk}(s) U_k(s) + V_k(s) \qquad k = 1, 2, \ldots, n \tag{5.10}$$

Equations (5.4) to (5.10) completely describe the control system with many variables. Figure 5.4 is a block diagram for a system with three engine outputs $Y_1(s)$, $Y_2(s)$, and $Y_3(s)$ and two controlled inputs $W_4(s)$ and $W_5(s)$. The entire system is enclosed except for the settings and the outside disturbances, which can be imposed on the system.

By eliminating $U_k(s)$, $Z_\nu(s)$, and $\Upsilon_\mu(s)$ from the previous system of equations, we have

$$Y_j(s) = \sum_{k=1}^{n} \left\{ \sum_{\nu=1}^{i} E_{jk}(s)S_{kk}(s)C_{k\nu}(s)[X_\nu(s) - L_{\nu\nu}(s)Y_\nu(s)] \right.$$
$$\left. + \sum_{\mu=i+1}^{n} E_{jk}(s)S_{kk}(s)C'_{k\mu}(s)[\Xi_\mu(s) - L_{\mu\mu}(s)W_\mu(s)] + E_{jk}V_k(s) \right\} \quad (5.11)$$

and

$$W_k(s) = \sum_{\nu=1}^{i} S_{kk}(s)C_{k\nu}(s)[X_\nu(s) - L_{\nu\nu}(s)Y_\nu(s)]$$
$$+ \sum_{\mu=i+1}^{n} S_{kk}(s)C'_{k\mu}(s)[\Xi_\mu(s) - L_{\mu\mu}W_\mu] + V_k \quad (5.12)$$

Equations (5.11) and (5.12) suggest a more compact block diagram for the system than Fig. 5.4. This, as shown in Fig. 5.5, involves a single

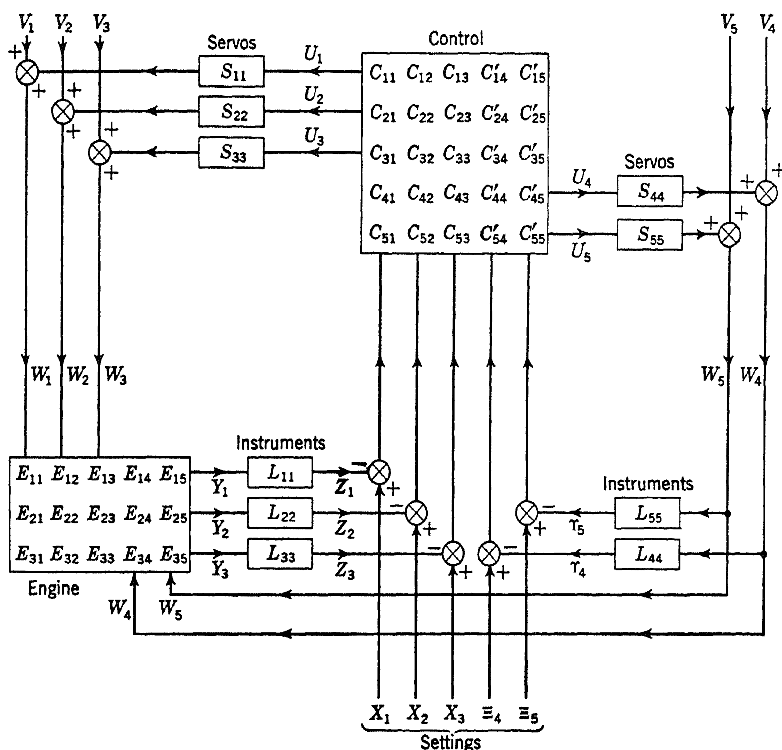


FIG. 5.4

systems matrix whose inputs are the error signals of the controlled variables and whose outputs are the controlled variables. The ESC matrix in Fig. 5.5 is a matrix in which the element in the jth row and νth column is $E_{jk}S_{kk}C_{k\nu}$. Similarly, for the ESC' matrix, the element in the jth row and μth column is $E_{jk}S_{kk}C'_{k\mu}$. Similarly, the elements of the SC matrix are $S_{kk}C_{k\nu}$ and the elements of the SC' matrix are $S_{kk}C'_{k\mu}$. The outside disturbances V_k are introduced through another matrix composed mainly of the engine matrix E.

5.3 Noninteraction Conditions. The criterion of noninteraction of controls can now be formulated in concrete terms. The problem is to

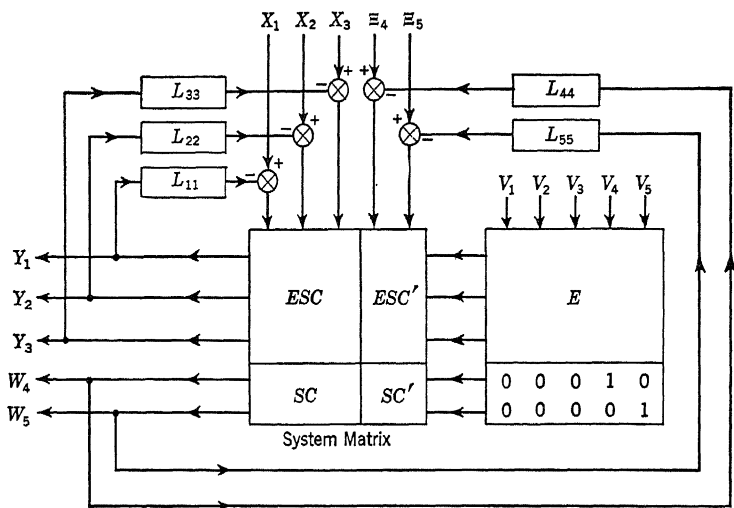


Fig. 5.5

determine conditions on the elements of the control matrix $C_{k\nu}(s)$ and $C'_{k\mu}(s)$ such that the settings $X_j(s)$ and $\Xi_\mu(s)$ will affect only their respective corresponding variables $Y_j(s)$ and $W_\mu(s)$, where $j = 1, 2, \ldots, i$ and $\mu = i + 1, \ldots, n$, and nothing else. Thus, for example, the setting $X_2(s)$ will modify only $Y_2(s)$, while the setting $\Xi_{i+1}(s)$ will modify only $W_{i+1}(s)$. The mathematical problem is thus one of "diagonalizing" the system matrix of Fig. 5.5. We put the design condition on the control matrix, because this is the part of the whole system most easily modified by the designer. The characteristics of the "engine," the servos, and the instruments are considered to be fixed and not at the disposal of the control engineer.

Let us study first a specific output $Y_g(s)$, with g assuming any one of the possible values $1, 2, \ldots, i$. Equations (5.11) and (5.12) can be written then as

$$Y_j(s) = \sum_{k=1}^{n} \Big[\sum_{\nu=1,\nu\neq g}^{i} E_{jk}S_{kk}C_{k\nu}(X_\nu - L_{\nu\nu}Y_\nu) + \sum_{\mu=i+1}^{n} E_{jk}S_{kk}C'_{k\mu}(\Xi_\mu - L_{\mu\mu}W_\mu) + E_{jk}V_k \Big] + \sum_{k=1}^{n} E_{jk}S_{kk}C_{kg}(X_g - L_{gg}Y_g)$$

and

$$W_k(s) = \sum_{\nu=1,\nu\neq g}^{i} S_{kk}C_{k\nu}(X_\nu - L_{\nu\nu}Y_\nu) + \sum_{\mu=i+1}^{n} S_{kk}C'_{k\mu}(\Xi_\mu - L_{\mu\mu}W_\mu) + V_k + S_{kk}C_{kg}(X_g - L_{gg}Y_g)$$

Now in order that a setting X_g will not influence any Y_j or W_μ except Y_g, the last terms of the above two equations must be zero for $j \neq g$ and for $k > i$. Therefore for any g among the set 1, 2, . . . , i,

$$\sum_{k=1}^{n} E_{jk}S_{kk}C_{kg} = 0 \qquad \text{for } j \neq g \tag{5.13}$$

and

$$C_{kg} = 0 \qquad \text{for } k > i \tag{5.14}$$

Equation (5.14) gives an immediate simplification of our control matrix. For instance, the example of Fig. 5.4 corresponds to $i = 3$, $n = 5$. Then Eq. (5.14) specifies that

$$C_{41} = C_{42} = C_{43} = C_{51} = C_{52} = C_{53} = 0$$

Equation (5.14) can also be used to simplify Eq. (5.13); it is in fact equivalent to

$$\sum_{k=1}^{i} E_{jk}S_{kk}C_{kg} = \sum_{k=1}^{i} \delta_{jg}E_{gk}S_{kk}C_{kg} \tag{5.15}$$

where g is any among the set 1, 2, . . . , i and δ_{jg} is the Kronecker delta, *i.e.*,

$$\left.\begin{aligned} \delta_{jg} &= 0 \qquad j \neq g \\ \delta_{jg} &= 1 \qquad j = g \end{aligned}\right\} \tag{5.16}$$

For any specific g, Eq. (5.15) is essentially a system of $i - 1$ linear algebraic equations for i unknowns $S_{kk}C_{kg}$, where $k = 1, 2, \ldots, i$. Therefore we can determine only the ratios of these unknowns but not their absolute values. This is exactly what is desired, as we do not wish to fix the control transfer function absolutely and thus lose freedom of design.

To find these ratios of the control transfer functions, we shall utilize a property of determinants: Let $|E^*_{jl}|$ be the cofactor of the E_{jl} element

in the determinant $|E^*|$ formed out of the square matrix E^*, then

$$\sum_{j=1}^{i} E_{jk}|E_{jl}^*| = 0 \qquad k \neq l$$

and (5.17)

$$\sum_{j=1}^{i} E_{jk}|E_{jl}^*| = |E^*| \qquad k = l$$

Multiplying Eq. (5.15) by $|E_{jl}^*|$ and summing over j, we obtain

$$\sum_{k=1}^{i} \sum_{j=1}^{i} |E_{jl}^*|\delta_{jg}E_{gk}S_{kk}C_{kg} = \sum_{k=1}^{i} \sum_{j=1}^{i} |E_{jl}^*|E_{jk}S_{kk}C_{kg}$$

Therefore, because of Eq. (5.17), we have

$$S_{ll}C_{lg} = |E_{gl}^*| \sum_{k=1}^{i} E_{gk}S_{kk}C_{kg} \Big/ |E^*| \qquad l = 1, 2, \ldots, i \tag{5.18}$$

In particular,

$$S_{gg}C_{gg} = |E_{gg}^*| \sum_{k=1}^{i} E_{gk}S_{kk}C_{kg} \Big/ |E^*|$$

Then by taking the ratio of Eq. (5.18) to the above equation, we can write

$$\frac{S_{jj}C_{j\nu}}{S_{\nu\nu}C_{\nu\nu}} = \frac{|E_{\nu j}^*|}{|E_{\nu\nu}^*|} \qquad j, \nu = 1, 2, \ldots, i \tag{5.19}$$

This equation gives the off-diagonal elements of the matrix SC in terms of the diagonal elements.

The conditions of Eqs. (5.14) and (5.19) are then the necessary conditions for noninteraction of the controlled variables Y_g. They were given by Boksenbom and Hood. The same authors proved that these conditions are also the sufficient conditions for noninteraction. Therefore the problem of finding the appropriate control matrix C is completely solved.

To solve the problem for the other part of the control matrix C', we have to consider the noninteraction conditions for the controlled variables W_μ, where $\mu = i + 1, \ldots, n$. For this purpose, we rewrite Eqs. (5.11) and (5.12) as

$$\begin{aligned} Y_j(s) = \sum_{k=1}^{n} \Big[\sum_{\nu=1}^{i} & E_{jk}S_{kk}C_{k\nu}(X_\nu - L_{\nu\nu}Y_\nu) \\ & + \sum_{\substack{\mu=i+1 \\ \mu\neq r}}^{n} E_{jk}S_{kk}C'_{k\mu}(\Xi_\mu - L_{\mu\mu}W_\mu) + E_{jk}V_k \Big] \\ & + \sum_{k=1}^{n} E_{jk}S_{kk}C'_{kr}(\Xi_r - L_{rr}W_r) \end{aligned} \tag{5.20}$$

and

$$W_k(s) = \sum_{\nu=1}^{i} S_{kk}C_{k\nu}(X_\nu - L_{\nu\nu}Y_\nu) + \sum_{\substack{\mu=i+1 \\ \mu \neq r}}^{n} S_{kk}C'_{k\mu}(\Xi_\mu - L_{\mu\mu}W_\mu) + V_k$$
$$+ S_{kk}C'_{kr}(\Xi_r - L_{rr}W_r) \quad (5.21)$$

where r is any among the set $i + 1, \ldots, n$ and $j = 1, 2, \ldots, i$. For the present purpose, the k index in Eq. (5.21) is any among the set $i + 1, \ldots, n$, because only these W_k's are the controlled variables. It is evident from Eqs. (5.20) and (5.21) that in order for the setting Ξ_r to influence only the variable W_r, the last terms in these equation must be zero. That is,

$$\sum_{k=1}^{n} E_{jk}S_{kk}C'_{kr} = 0 \qquad j = 1, 2, \ldots, i \qquad (5.22)$$

and

$$C'_{kr} = 0 \qquad \text{for } k, r = i + 1, \ldots, n \text{ and } k \neq r \qquad (5.23)$$

Again, Eq. (5.23) gives an immediate simplification of the control matrix: For the example represented by Fig. 5.4, $i = 3$, $n = 5$, and so

$$C'_{45} = C'_{54} = 0$$

Equation (5.23) can be used also to simplify Eq. (5.22). That equation is reduced to

$$\sum_{k=1}^{i} E_{jk}S_{kk}C'_{kr} = -E_{jr}S_{rr}C'_{rr}$$

Multiply both sides of the above equation by $|E^*_{jl}|$ and sum over j. Then

$$\sum_{k=1}^{i} \sum_{j=1}^{i} |E^*_{jl}|E_{jk}S_{kk}C'_{kr} = -S_{rr}C'_{rr} \sum_{j=1}^{i} |E^*_{jl}|E_{jr}$$

By using the properties of determinants as given by Eq. (5.17), we have

$$|E^*|S_{ll}C'_{lr} = -S_{rr}C'_{rr} \sum_{j=1}^{i} |E^*_{jl}|E_{jr}$$

Therefore we can write the above equation in the following form, by replacing l by j, and j by l,

$$\frac{S_{jj}C'_{jr}}{S_{rr}C'_{rr}} = -\frac{1}{|E^*|} \sum_{l=1}^{i} |E^*_{lj}|E_{lr} \qquad \begin{matrix} j = 1, \ldots, i \\ r = i + 1, \ldots, n \end{matrix} \qquad (5.24)$$

This equation then gives the off-diagonal elements of the control matrix SC' in terms of the diagonal elements. Equations (5.23) and (5.24) are

the necessary and sufficient conditions of noninteraction for the controlled variables $W_\mu(s)$, for $\mu = i + 1, \ldots, n$.

For complete noninteraction of all controlled variables, the conditions specified by Eqs. (5.14), (5.19), (5.23), and (5.24) must be satisfied. The off-diagonal elements of the complete control matrix are either zero or expressed in terms of the diagonal elements. When the engine characteristics as expressed by the engine matrix are known, the diagonal elements of the control matrix determine completely the whole control matrix.

5.4 Response Equations. With the noninteraction conditions all satisfied, Eqs. (5.11) and (5.12) can be made much simpler. For instance, by interchanging the two summations,

$$Y_j(s) = \sum_{\nu=1}^{i} [X_\nu(s) - L_{\nu\nu}(s) Y_\nu(s)] \sum_{k=1}^{n} E_{jk} S_{kk} C_{k\nu} + \sum_{\mu=i+1}^{n} [\Xi_\mu(s) - L_{\mu\mu}(s) W_\mu(s)] \sum_{k=1}^{n} E_{jk} S_{kk} C'_{k\mu} + \sum_{k=1}^{n} E_{jk} V_k$$

But according to Eqs. (5.13) and (5.14), the sum over k of the first term vanishes except when $\nu = j$. According to Eq. (5.22), the second term vanishes. Thus

$$Y_j(s) = [X_j(s) - L_{jj}(s) Y_j(s)] \sum_{k=1}^{i} E_{jk} S_{kk} C_{kj} + \sum_{k=1}^{n} E_{jk} V_k$$

Now $S_{kk}C_{kj}$ can be expressed in terms of the diagonal element $S_{jj}C_{jj}$ according to Eq. (5.19). Thus, using Eq. (5.17), we have

$$\sum_{k=1}^{i} E_{jk} S_{kk} C_{kj} = \frac{S_{jj}C_{jj}}{|E^*_{jj}|} \sum_{k=1}^{i} E_{jk} |E^*_{jk}| = S_{jj} C_{jj} \frac{|E^*|}{|E^*_{jj}|}$$

Therefore, finally,

$$Y_j(s) = \frac{|E^*|}{|E^*_{jj}|} S_{jj} C_{jj} [X_j(s) - L_{jj}(s) Y_j(s)] + \sum_{k=1}^{n} E_{jk} V_k \tag{5.25}$$

By using the noninteraction conditions, a similar calculation will reduce Eq. (5.12) to

$$W_\mu(s) = S_{\mu\mu} C'_{\mu\mu} [\Xi_\mu(s) - L_{\mu\mu}(s) W_\mu(s)] + V_\mu(s) \qquad \mu = i + 1, \ldots, n \tag{5.26}$$

By writing

$$R_{jj} = \frac{|E^*| S_{jj} C_{jj}}{|E^*| S_{jj} C_{jj} L_{jj} + |E^*_{jj}|} \tag{5.27}$$

and

$$R'_{\mu\mu} = \frac{S_{\mu\mu}C'_{\mu\mu}}{S_{\mu\mu}C'_{\mu\mu}L_{\mu\mu} + 1} \tag{5.28}$$

the solutions of Eqs. (5.25) and (5.26) can be written as

$$Y_j(s) = R_{jj}(s)X_j(s) - [R_{jj}(s)L_{jj}(s) - 1] \sum_{k=1}^{n} E_{jk}(s)V_k(s) \tag{5.29}$$

and

$$W_\mu(s) = R'_{\mu\mu}(s)\Xi_\mu(s) - [R'_{\mu\mu}(s)L_{\mu\mu}(s) - 1]V_\mu(s) \tag{5.30}$$

Equations (5.29) and (5.30) give the relations for calculating the controlled variables from the settings and the disturbances. They are quite similar to Eq. (5.3) for the simple system of one controlled variable. The function $R_{jj}(s)$ is the over-all transfer function from input $X_j(s)$ to output $Y_j(s)$. The function $R'_{\mu\mu}(s)$ is the over-all transfer function from the input $\Xi_\mu(s)$ to output $W_\mu(s)$. These over-all transfer functions are calculated according to Eqs. (5.27) and (5.28) using the characteristics of the engine, the servos, the instruments, and the control. In fact, the procedure of design will be to determine for each j and μ the proper control transfer function $C_{jj}(s)$ or $C'_{\mu\mu}(s)$ for satisfactory performance by methods explained in Chap. 4. The nondiagonal elements of the control matrix are then determined by Eqs. (5.14), (5.19), (5.23), and (5.24). When this is done, we have a noninteracting control of good performance for a complex, many-variable system.

5.5 Turbopropeller Control. As a simple example of the general theory of noninteracting controls, let us consider the control problem of a turbopropeller engine (Fig. 5.6). The variables of the operation of such an engine are the speed of rotation, the turbine-inlet temperature, the propeller blade angle, and the fuel rate. The control system has to be designed for various possible steady-state normal operating conditions. For each steady-state operating condition, we have to investigate the control performance in nonsteady states near that particular operating point. Let $W_1(s)$ be the Laplace transform of the deviation of propeller blade angle from the normal point, and $W_2(s)$ be the Laplace transform of the deviation of fuel rate from the normal value. Then since we are interested in nonsteady states *near* the normal point, the relation between the excess of turbine torque over the torque absorbed by the compressor and the propeller, and the propeller blade angle and fuel rate can be linearized. Therefore the excess torque is represented by a linear combination of $W_1(s)$ and $W_2(s)$. Let the Laplace transform of the deviation of the rotating speed from its normal value be $Y_1(s)$. Then the excess torque is represented by $(1 + \tau s)Y_1(s)$, where τ is the characteristic time constant due to inertia of the rotating components of the power plant,

cf. Eq. (4.1). The value of τ depends upon the normal operating point. Thus

$$(1 + \tau s)Y_1(s) = -aW_1(s) + bW_2(s) \tag{5.31}$$

where a and b are positive real constants, deduced from the engine characteristics near the normal operating point. The physical meanings of a and b are as follows: If the fuel rate is held at the normal value, $W_2(s) \equiv 0$. Equation (5.31) then gives a as $-Y_1(0)/W_1(0)$. But $s = 0$ corresponds to the steady state, and therefore a is the ratio of *decrease* of the steady-state engine speed to increase in propeller-blade angle with the fuel rate held constant. If the steady-state engine performance is given as graphs of engine speed versus propeller blade angle at various constant fuel rates, then a is the slope of this graph evaluated at the chosen normal operating point. Similarly, the meaning of b is the slope of the

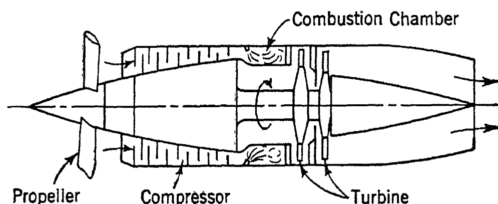


FIG. 5.6

steady-state engine-speed versus fuel-rate curve with constant propeller blade angle, evaluated at the chosen normal operating point. Thus the constants a and b are specified by the steady-state performance curves of the engine.

For an axial compressor, the mass air flow through the compressor for a certain inlet condition is almost constant at a given compressor speed. Therefore, with given inlet conditions, the ratio of heat added to the gas to the mass of the gas is a function of engine speed and fuel rate. Hence the engine speed and fuel rate determine the turbine-inlet temperature. Let $Y_2(s)$ be the Laplace transform of the deviation of turbine-inlet temperature from the normal value. Then an equation between $Y_2(s)$ and $W_1(s)$, with $W_2(s)$ similar to that in Eq. (5.31), can be established. However, since the characteristic time for reaching thermal equilibrium of the gas is practically zero, the relation is simpler:

$$Y_2(s) = cW_2(s) - eY_1(s) \tag{5.32}$$

where c and e are again positive real constants. In fact, c is the slope of the turbine-inlet temperature versus fuel-rate curve at constant engine speed, while e is the slope of the turbine-inlet temperature versus engine-speed curve at constant fuel rate; all evaluated at the chosen steady-state operating point.

By solving for $Y_1(s)$ and $Y_2(s)$ in Eqs. (5.31) and (5.32), we have

$$\left.\begin{aligned} Y_1(s) &= \frac{-a}{1+\tau s} W_1(s) + \frac{b}{1+\tau s} W_2(s) \\ Y_2(s) &= \frac{ae}{1+\tau s} W_1(s) + \frac{(c-be)+c\tau s}{1+\tau s} W_2(s) \end{aligned}\right\} \tag{5.33}$$

These equations specify the engine matrix E in our theory. It is interesting to note that there is only one time constant τ in the engine matrix. Only this time characteristic is intrinsic to the engine. The complete control system has, of course, other time constants. But the other time constants are introduced by the control functions, the servos, and the instruments, and are not in the engine matrix.

Let us consider first the case of controlling the engine speed and the fuel rate. Thus the controlled variables are $Y_1(s)$ and $W_2(s)$. We need, then, only the first equation of Eq. (5.33), and $i = 1$ and $n = 2$. Hence the engine matrix has only two elements:

$$E_{11} = \frac{-a}{1+\tau s} \qquad E_{12} = \frac{b}{1+\tau s} \tag{5.34}$$

and

$$|E^*| = |E^*_{11}|E_{11} = E_{11} \qquad |E^*_{11}| = 1 \tag{5.35}$$

The control system is represented by

$$\begin{aligned} U_1(s) &= C_{11}(s)[X_1(s) - L_{11}(s)Y_1(s)] + C'_{12}(s)[\Xi_2(s) - L_{22}(s)W_2(s)] \\ U_2(s) &= C_{21}(s)[X_1(s) - L_{11}(s)Y_1(s)] + C'_{22}(s)[\Xi_2(s) - L_{22}(s)W_2(s)] \end{aligned} \tag{5.36}$$

The noninteraction conditions require

$$C_{21}(s) = 0 \tag{5.37}$$

and, using Eq. (5.35),

$$\frac{S_{11}(s)C'_{12}(s)}{S_{22}(s)C'_{22}(s)} = \frac{-|E^*_{11}|E_{12}}{|E^*|} = -\frac{E_{12}}{E_{11}} = \frac{b}{a} \tag{5.38}$$

Since $-a$ is the partial derivative of engine speed with respect to the propeller blade angle, and b is the partial derivative of engine speed with respect to the fuel rate, the ratio b/a is the rate of change of the propeller blade angle with change in fuel rate at constant engine speed. Clearly this ratio is a function of flight conditions of the turbopropeller engine. For instance, this ratio b/a increases as the altitude increases. Therefore a properly designed control requires means of compensating for changes in flight and operating conditions of the engine.

The response function $R_{11}(s)$ for the engine speed is then

$$R_{11}(s) = \frac{aS_{11}(s)C_{11}(s)}{aS_{11}(s)C_{11}(s)L_{11}(s) - (1 + \tau s)}$$

and for the fuel rate

$$R'_{22}(s) = \frac{S_{22}(s)C'_{22}(s)}{S_{22}(s)C'_{22}(s)L_{22}(s) + 1} \tag{5.39}$$

These equations determine the noninteracting response characteristics of the engine speed and the fuel rate. The problem is to design the control transfer functions $C_{11}(s)$ and $C'_{22}(s)$ in such a way that the performance is satisfactory for the full range of expected operating conditions.

Now consider the second possibility of turbopropeller control. We shall control the engine speed and the turbine-inlet temperature. The controlled variables are then $Y_1(s)$ and $Y_2(s)$. Thus in this case, we need both equations of Eq. (5.33), and $i = n = 2$. Then the noninteraction conditions specify

$$\frac{S_{22}(s)C_{21}(s)}{S_{11}(s)C_{11}(s)} = -\frac{ae}{(c - be) + c\tau s}$$

and

$$\frac{S_{11}(s)C_{12}(s)}{S_{22}(s)C_{22}(s)} = \frac{b}{a} \tag{5.40}$$

The response function for the engine speed is then

$$R_{11}(s) = \frac{S_{11}(s)C_{11}(s)}{S_{11}(s)C_{11}(s)L_{11}(s) - \dfrac{(c - be) + c\tau s}{ac}} \tag{5.41}$$

and for the turbine-inlet temperature

$$R_{22}(s) = \frac{S_{22}(s)C_{22}(s)}{S_{22}(s)C_{22}(s)L_{22}(s) + (1/c)}$$

5.6 Turbojet Engine with Afterburning. We shall now treat the problem of controlling a turbojet engine with afterburning, mentioned at the beginning of this chapter. The physical components are sketched in Fig. 5.7. We shall again study the problem of control for nonsteady states *near* a chosen normal steady-state operating point. Therefore linearization of the relations between the different variables is allowed.

Let $Y_1(s)$ again denote the Laplace transform of the deviation of engine speed from the normal value, $W_1(s)$ the Laplace transform of the deviation of the tail-pipe opening from the normal value, $W_2(s)$ the Laplace transform of the deviation of combustion-chamber fuel rate from the normal, and, finally, $W_3(s)$ the Laplace transform of the deviation of

the tail-pipe fuel rate from the normal. Then we can write, in a form similar to Eq. (5.31) for the turbopropeller.

$$(1 + \tau s)Y_1(s) = a_1W_1(s) + a_2W_2(s) + a_3W_3(s) \tag{5.42}$$

where a_1, a_2, and a_3 are real constants. As in the case of the turbopropeller, these constants are slopes of the steady-state performance curves of the engine. Thus a_1 is the rate of change of engine speed with respect to tail-pipe opening at constant fuel rates to the engine combustion chamber and to the tail pipe. a_2 is the rate of change of engine speed with respect to the engine fuel rate. a_3 is the rate of change of engine speed with respect to the tail-pipe fuel rate. τ in

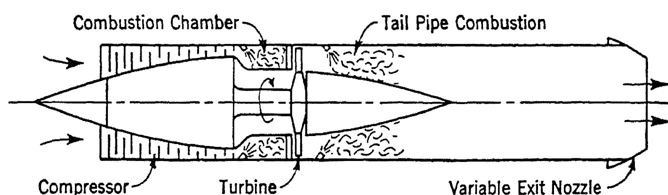


Fig. 5.7

Eq. (5.42) is again the only characteristic time of the engine system and represents the effects of the inertia of the rotating components. This linearized relation between the engine speed and other engine-input variables was derived by M. S. Feder and R. Hood.[1]

If the compressor of the engine is an axial compressor, Eq. (5.32) of the previous section is again applicable here. $Y_2(s)$ represents the turbine-inlet temperature, and thus

$$Y_2(s) = -eY_1(s) + cW_2(s)$$

By solving for $Y_1(s)$ and $Y_2(s)$ in the above equation and in Eq. (5.42), we have

$$\left.\begin{aligned} Y_1(s) &= \frac{a_1}{1 + \tau s} W_1(s) + \frac{a_2}{1 + \tau s} W_2(s) + \frac{a_3}{1 + \tau s} W_3(s) \\ Y_2(s) &= -\frac{a_1 e}{1 + \tau s} W_1(s) + \frac{(c - a_2 e) + c\tau s}{1 + \tau s} W_2(s) \\ &\qquad - \frac{a_3 e}{1 + \tau s} W_3(s) \end{aligned}\right\} \tag{5.43}$$

The elements of the engine matrix are thus

$$\left.\begin{aligned} E_{11} &= \frac{a_1}{1 + \tau s} & E_{12} &= \frac{a_2}{1 + \tau s} & E_{13} &= \frac{a_3}{1 + \tau s} \\ E_{21} &= -\frac{a_1 e}{1 + \tau s} & E_{22} &= \frac{(c - a_2 e) + c\tau s}{1 + \tau s} & E_{23} &= -\frac{a_3 e}{1 + \tau s} \end{aligned}\right\} \tag{5.44}$$

[1] M. S. Feder and R. Hood, *NACA TN* 2183 (1950).

Let us consider the problem of controlling the engine speed, the turbine-inlet temperature, and the tail-pipe fuel rate. The controlled variables are then $Y_1(s)$, $Y_2(s)$, and $W_3(s)$. The control equations are then

$$\left.\begin{aligned}
U_1(s) &= C_{11}(s)[X_1(s) - L_{11}(s)Y_1(s)] \\
&\quad + C_{12}(s)[X_2(s) - L_{22}(s)Y_2(s)] + C'_{13}(s)[\Xi_3(s) - L_{33}(s)W_3(s)] \\
U_2(s) &= C_{21}(s)[X_1(s) - L_{11}(s)Y_1(s)] \\
&\quad + C_{22}(s)[X_2(s) - L_{22}(s)Y_2(s)] + C'_{23}(s)[\Xi_3(s) - L_{33}(s)W_3(s)] \\
U_3(s) &= C_{31}(s)[X_1(s) - L_{11}(s)Y_1(s)] \\
&\quad + C_{32}(s)[X_2(s) - L_{22}(s)Y_2(s)] + C'_{33}(s)[\Xi_3(s) - L_{33}(s)W_3(s)]
\end{aligned}\right\} \tag{5.45}$$

where $X_1(s)$, $X_2(s)$, and $\Xi_3(s)$ are the settings for the engine speed, the turbine-inlet temperature, and the tail-pipe fuel rate, respectively.

The noninteraction condition of Eq. (5.14) requires immediately that

$$C_{31}(s) = C_{32}(s) = 0 \tag{5.46}$$

The condition of Eq. (5.19) gives

$$\left.\begin{aligned}
\frac{S_{11}(s)C_{12}(s)}{S_{22}(s)C_{22}(s)} &= -\frac{a_2}{a_1} \\
\frac{S_{22}(s)C_{21}(s)}{S_{11}(s)C_{11}(s)} &= \frac{a_1 e}{(c - a_2 e) + c\tau s}
\end{aligned}\right\} \tag{5.47}$$

The noninteraction condition of Eq. (5.24) gives

$$\frac{S_{11}(s)C'_{13}(s)}{S_{33}(s)C'_{33}(s)} = -\frac{a_3}{a_1}$$

and

$$C'_{23}(s) = 0 \tag{5.48}$$

The ratios $-a_2/a_1$ and $-a_3/a_1$ in the above equations have simple physical meanings: $-a_2/a_1$ is the rate of change of tail-pipe opening with respect to engine fuel rate at constant engine speed and constant tail-pipe fuel rate. $-a_3/a_1$ is the rate of change of tail-pipe opening with respect to tail-pipe fuel rate at constant engine speed and constant engine fuel rate.

When Eqs. (5.46) to (5.48) are satisfied, we have noninteracting control, and the response function for the engine speed is

$$R_{11}(s) = \frac{S_{11}(c)C_{11}(s)}{S_{11}(s)C_{11}(s)L_{11}(s) + \dfrac{(c - a_2 e) + c\tau s}{a_1 c}} \tag{5.49}$$

The response function for the turbine-inlet temperature is

$$R_{22}(s) = \frac{S_{22}(s)C_{22}(s)}{S_{22}(s)C_{22}(s)L_{22}(s) + (1/c)} \tag{5.50}$$

The response function for the tail-pipe fuel rate is

$$R'_{33}(s) = \frac{S_{33}(s)C'_{33}(s)}{S_{33}(s)C'_{33}(s)L_{33}(s) + 1} \tag{5.51}$$

These equations then give the starting point of proper design of the control transfer functions $C_{11}(s)$, $C_{22}(s)$, $C'_{33}(s)$, and hence $C_{12}(s)$, $C_{21}(s)$, and $C'_{13}(s)$.

CHAPTER 6

ALTERNATING-CURRENT SERVOMECHANISMS AND OSCILLATING CONTROL SERVOMECHANISMS

In this chapter and the two following chapters, we shall extend the concepts and methods developed for simple servomechanisms in Chaps. 2 and 3 to linear systems which are more complicated but nevertheless can be treated by *approximately* the same technique. Therefore they demonstrate the power of the basic principles of servomechanism design. The contents of this and the next chapter follow closely the treatment of L. A. MacColl.[1]

6.1 Alternating-current Systems. So far, whenever we have been considering a servomechanism containing an electric motor, we have assumed implicitly that the motor is a d-c motor. In practice, however, it may well be desirable to use a-c motors. It is clear that the use of such motors necessitates the reconsideration of some parts of our previous discussion.

Consider a servomechanism as sketched in Fig. 6.1. The purpose of the system is to turn the motor to angle ϕ, according to the input signal. The output angle ϕ is measured by a potentiometer. The voltage across the potentiometer is the feedback signal. In this system all of the currents and voltages appearing in the amplifier, motor, and potentiometer are modulated sinusoids, *i.e.*, sinusoidal functions of fixed frequency, say ω_0, but with time-varying amplitude. The basic alternating current is generated by the oscillator. When a certain condition, which will be discussed presently, is satisfied, much of the earlier theory is applicable to this system.

Let us consider for a moment the general steady-state theory of linear systems of constant coefficients subjected to signals which are modulated sinusoids. Here the expression "steady state" refers to the fact that the modulating signals are assumed to be purely sinusoidal functions of time. Let the unmodulated "carrier" be $\cos \omega_0 t$. The phase angle is here neglected without loss of generality. Since the carrier is expressed in real form, it is obviously legitimate to take the

[1] L. A. MacColl, "Fundamental Theory of Servomechanisms," D. Van Nostrand Company, Inc., New York, 1945.

modulating signal in complex form, $e^{i\omega t}$. Then the modulated carrier is

$$x(t) = e^{i\omega t} \cos \omega_0 t = \tfrac{1}{2}[e^{i(\omega+\omega_0)t} + e^{i(\omega-\omega_0)t}] \tag{6.1}$$

and the steady-state response $y_{\text{steady}}(t)$ of a system having the transfer function $F(s)$ is, according to Eq. (2.16),

$$y_{\text{steady}}(t) = \tfrac{1}{2}[F(i\omega + i\omega_0)e^{i\omega_0 t} + F(i\omega - i\omega_0)e^{-i\omega_0 t}]e^{i\omega t} \tag{6.2}$$

For a physical system, the function $F(s)$ is generally a ratio of polynomials in s with real coefficients. Then, as has already been indicated

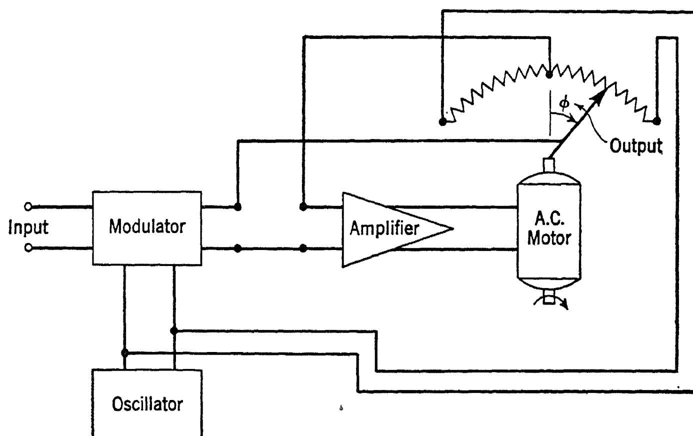


FIG. 6.1

by Eq. (3.17),

$$F(-i\omega) = \overline{F(i\omega)} \tag{6.3}$$

where the bar over the symbol indicates the complex conjugate value. Therefore we can write Eq. (6.2) as

$$\tfrac{1}{2}[F^*(i\omega)e^{i\omega_0 t} + F^*(\overline{-i\omega})e^{-i\omega_0 t}]e^{i\omega t} \tag{6.4}$$

where

$$F^*(i\omega) = F(i\omega + i\omega_0) \tag{6.5}$$

Now we suppose that the system is such that we have the relation

$$F(i\omega_0 + i\omega) = \overline{F(i\omega_0 - i\omega)} \tag{6.6}$$

Then the expression of (6.4) can be written in the form

$$F^*(i\omega)e^{i\omega t} \cos \omega_0 t$$

This result shows that when the condition of (6.5) is satisfied, the amplitude of the response of the system to the modulated carrier of Eq. (6.1) is the same as the response of a system having the frequency response

$F^*(i\omega)$ at the input frequency ω. This statement can be immediately generalized to apply to more general input functions by the principle of superposition for linear systems. If Eq. (6.6) is satisfied, at least approximately, throughout a range of values of ω which includes the more important parts of the Fourier spectrum of a modulating input signal $x(t)$, the amplitude of the resulting modulated output signal is at least approximately equal to the response of a system having the frequency response $F^*(i\omega)$ to the input signal $x(t)$. We have shown in Chap. 4 that the performance of the feedback servomechanism is determined completely by the frequency response. The approximate frequency response now is $F^*(i\omega)$. Then all the methods for determining the performance of a system developed in Chap. 4 can be applied to the a-c systems. The only difference is the use of $F^*(i\omega)$ instead of $F(i\omega)$ in the analysis.

6.2 Translation of the Transfer Function to a Higher Frequency. If we leave out of account certain trivial systems, *e.g.*, pure resistances, it follows from Eq. (6.3) that

$$F(i\omega_0 + i\omega) = \overline{F(-i\omega_0 - i\omega)}$$

This is different from the condition of Eq. (6.6). Therefore Eq. (6.6) cannot be satisfied exactly for all real values of ω. Or, if we alter our point of view slightly, we can say that the frequency responses $F(i\omega)$ and $F^*(i\omega)$ of two *physical* systems cannot rigorously satisfy the relation of Eq. (6.5) for all real values of ω. Nevertheless, it is entirely possible, and indeed quite common, for the frequency responses $F^*(i\omega)$ and $F(i\omega)$ of two physical systems to satisfy the relation of Eq. (6.5) approximately over a range of values of ω which is large enough to include the more important parts of the Fourier spectra of the particular input signals we are concerned with. We can see this briefly as follows:

Consider the impedance Z of an inductance L and a capacitance C in series, at a frequency ω'

$$\begin{aligned} Z &= Li\omega' + \frac{1}{Ci\omega'} \\ &= Li\omega'\left(1 - \frac{1}{LC\omega'^2}\right) \end{aligned}$$

If we make L and C of such magnitude that

$$\omega_0^2 = \frac{1}{LC} \tag{6.7}$$

then

$$Z = Li\omega'\left(1 - \frac{\omega_0^2}{\omega'^2}\right) = Li(\omega' - \omega_0)\left(1 + \frac{\omega_0}{\omega'}\right)$$

For $\omega' - \omega_0 = \omega$ small, or $\omega + \omega_0$ near ω_0,

$$Z \approx 2Li(\omega' - \omega_0) = 2Li\omega$$

That is, at the frequency $\omega' = \omega_0 + \omega$, the impedance of the series combination of L and C satisfying Eq. (6.7) is approximately equal to the impedance of an inductance $2L$ at the frequency ω.

Similarly, consider the impedance Z of an inductance L and a capacitance C in parallel, at the frequency ω'

$$\begin{aligned} \frac{1}{Z} &= \frac{1}{Li\omega'} + Ci\omega \\ &\approx 2Ci(\omega' - \omega_0) = 2Ci\omega \end{aligned}$$

if condition (6.7) is satisfied. Therefore at the frequency $\omega' = \omega + \omega_0$, the impedance of the parallel combination of L and C satisfying Eq. (6.7) is approximately equal to the impedance of a capacitance $2C$ at the frequency ω.

The impedance of a pure resistance is, of course, independent of the frequency—it is the same at the frequency $\omega + \omega_0$ as at the frequency ω. Thus, starting from a physical system with a transfer function $F^*(s)$, we can, by replacing any inductance L by a series combination of inductance $\frac{1}{2}L$ and capacity $C = 2/L\omega_0^2$ and by replacing any capacity C by a parallel combination of capacity $\frac{1}{2}C$ and inductance $L = 2/C\omega_0^2$, obtain a physical system having a transfer function $F(s)$ such that the relationship of Eq. (6.6) is satisfied approximately for small values of ω. This procedure of going from $F^*(s)$ to $F(s)$ is called translating the transfer function on the frequency scale by ω_0.

Let ω_0 denote the frequency of the current supplied by the oscillator. It is clear that all of the currents and voltages in the system are modulations of the carrier wave $\cos \omega_0 t$. Hence it immediately follows from the above that, in order to design the amplifier for the system with alternating current, we need only design a suitable amplifier for a system with direct current by methods described in Chap. 4 and then translate the characteristics of the amplifier upward on the ω scale by the amount ω_0, in accordance with the procedure sketched above.

As we have indicated, the foregoing arguments involve a considerable number of approximations of one kind or another. An entirely complete discussion of servomechanisms with alternating current would necessitate an examination of the effects of all these approximations. We shall not go into this investigation, because it would be involved and tedious, and because it does not appear to be a very urgent matter as far as servomechanism art is concerned.

6.3 Oscillating Control Servomechanisms. We shall now consider another class of systems, which we call oscillating control servomecha-

nisms. These resemble servomechanisms with a-c motors, in that in both cases the signals are caused to modulate a periodic oscillation. However, in the case of oscillating control servomechanisms the modulation employed is not the ordinary amplitude modulation. In order to introduce the concept of an oscillating control servomechanism intelligibly, we must first give a little preparatory discussion.

One very primitive but very common kind of servomechanism can be described as follows. Suppose that to the system we were to add a circuit containing a relay, designed to function so that no voltage would be applied to the output terminals unless the absolute value of the input $x(t)$ exceeded a certain threshold, and so that when $|x|$ did exceed the threshold the output would be the full electromotive force E of a source applied with such a polarity as to tend to reduce the absolute value of the error. We would then have an example of what we call an on-off servomechanism.

On-off servomechanisms have the great advantage that comparatively simple systems of this kind can be made to handle large amounts of power. This is often difficult to achieve with servomechanisms of other types. On the other hand, on-off servomechanisms are definitely nonlinear systems, and, as will be shown in Chap. 10, their performances tend to be inferior to those of the systems we have considered previously. Briefly, an oscillating control servomechanism is a modification of an on-off servomechanism, which enables us to secure the advantage of linearity without sacrificing the advantage of large power-carrying capacity.

Before proceeding to the treatment of oscillating control servomechanisms proper, we shall present a general theoretical result, upon which the theory of all such systems is based. Let us consider a device having the following property: According as the input signal $x(t)$ is positive or negative, the output signal $y(t)$ is $+A$ or $-A$, where A is a fixed constant. We may think of such a device as an ideal relay, an on-off system with zero threshold. Suppose that the input signal to the relay is

$$x(t) = E_0 \sin \omega_0 t + kE_0 \sin \omega t \tag{6.8}$$

where E_0, k, ω_0, and ω are constants. In connection with oscillating control servomechanisms, the term $E_0 \sin \omega_0 t$ will be a persistent oscillation in the system, and $kE_0 \sin \omega t$ will be an applied signal or modulating signal. We shall calculate the corresponding output $y(t)$ presently.

6.4 Frequency Response of a Relay. The output of the relay in response to the input of Eq. (6.8) can be written in the form

$$\sum_{m=0}^{\infty} \sum_{n=-\infty}^{\infty} a_{nm} \sin [(m\omega_0 + n\omega)t] \tag{6.9}$$

where the a's are independent of t. When $m = 0$, the inner summation is to be extended only over positive values of n. For our purposes the only coefficients that are of any immediate interest are a_{10} and a_{01}; for in the case of an oscillating control servomechanism operating under normal conditions, the other coefficients either are negligibly small, or they correspond to oscillations which are suppressed by suitable filtering.

When $k = 0$, *i.e.*, when the input to the relay is a purely sinusoidal function with frequency ω_0, the output from the relay is obviously a series of alternate positive and negative square waves of height A and

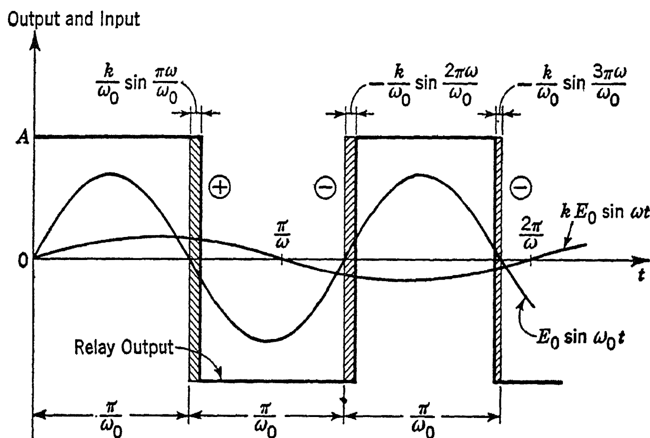


FIG. 6.2

duration $\omega_0/4\pi$. It is known that the amplitude a_{01} of the leading term of a Fourier expansion of such a square wave is equal to

$$a_{01} = \frac{4A}{\pi} \tag{6.10}$$

When $k \neq 0$, the output from the relay is presented in Fig. 6.2. The difference of the outputs with $k \neq 0$ and with $k = 0$ is a series of rectangles of height $2A$, indicated as shaded areas in Fig. 6.2. When $|k| \ll 1$, the change of switching points of the output from the evenly spaced points of $t_n = n\pi/\omega_0$ is very small. Thus the correction rectangles are very narrow, as shown in the figure. The width of these rectangles can be calculated as the value of the modulating signal at t_n divided by the slope of the persistent oscillation at t_n. Thus the width is

$$\left| \frac{kE_0 \sin \omega t_n}{E_0 \cos \omega_0 t_n} \right| = \frac{k}{\omega_0} |\sin \omega t_n|$$

These rectangular areas are to be added (+) or to be subtracted (−) from the unmodulated output according to whether $\sin \omega t_n$ is positive or

negative. Thus the areas of the rectangles can be considered as

$$\frac{2Ak}{\omega_0}\sin \omega t_n$$

The coefficient a_{10} in Eq. (6.9) is the coefficient of the leading $\sin \omega t$ term in the Fourier expansion of this series of narrow rectangular waves. Since the value of $\sin \omega t$ at the rectangular areas is $\sin \omega t_n$, we have, by taking N such correction rectangles,

$$a_{10}\int_0^{N\pi/\omega_0}\sin^2 \omega t\, dt = 2A\frac{k}{\omega_0}\sum_{n=0}^{N}\sin^2 \omega t_n$$

But

$$\begin{aligned}\int_0^{N\pi/\omega_0}\sin^2 \omega t\, dt &= \tfrac{1}{2}\int_0^{N\pi/\omega_0}(1-\cos 2\omega t)\, dt\\ &= \frac{1}{2}\frac{N\pi}{\omega_0} - \frac{1}{4\omega}\sin\left(2N\pi\frac{\omega}{\omega_0}\right)\end{aligned}$$

and

$$\begin{aligned}\sum_{n=0}^{N}\sin^2 \omega t_n = \sum_{n=1}^{N}\sin^2\left(n\pi\frac{\omega}{\omega_0}\right) &= \frac{1}{2}\sum_{n=1}^{N}\left[1-\cos\left(2n\pi\frac{\omega}{\omega_0}\right)\right]\\ &= \frac{N}{2} - \frac{1}{2}\sum_{n=1}^{N}\cos\left(2n\pi\frac{\omega}{\omega_0}\right)\end{aligned}$$

The sum remains finite as we increase N indefinitely. Therefore, by making N large, we have

$$a_{10} = 2\frac{Ak}{\pi} \tag{6.11}$$

Equations (6.10) and (6.11) give the two important coefficients a_{01} and a_{10} for small k. For general values of k, these coefficients were computed by R. M. Kalb and W. R. Bennett.[1] When $0 < k < 1$,

$$\left.\begin{aligned}a_{01} &= \frac{8A}{\pi^2}E(k)\\ a_{10} &= \frac{8A}{\pi^2 k}[E(k) - (1-k^2)K(k)]\end{aligned}\right\} \tag{6.12}$$

where $K(k)$ and $E(k)$ denote the complete elliptic integrals of the first and the second kinds, respectively. For k small, the elliptic integrals can be expanded; then

[1] *Bell System Tech. J.*, **14**, 322–359 (1935).

$$\left.\begin{aligned} a_{01} &= \frac{4A}{\pi}\left(1 - \frac{k^2}{4} - \cdots\right) \\ a_{10} &= \frac{2Ak}{\pi}\left(1 + \frac{k^2}{8} + \cdots\right) \end{aligned}\right\} \tag{6.13}$$

Equation (6.13) shows that our simple computation is correct within the accuracy of analysis. However, it also shows that the simple results of Eqs. (6.10) and (6.11) are accurate enough for even moderate values of k. Therefore the ratio of the component of frequency ω in the output to the component of the same frequency in the input, *i.e.*, the frequency response $F_l(i\omega)$, is approximately equal to

$$F_l(i\omega) = \frac{2A}{\pi E_0} \tag{6.14}$$

As shown by Eqs. (6.10) and (6.13), when k is small, the amplitude of the component of frequency ω_0 in the output is approximately a constant determined entirely by the properties of the relay. Also, the ratio of the component of frequency ω_0 in the output to the component of the same frequency in the input is $4A/(\pi E_0)$. Thus the amplification of the relay for the component of frequency ω_0 is 6 db greater than the amplification for the component of frequency ω.

Now it is easily seen that the preceding considerations can be extended to the case in which we have, instead of the signal $kE_0 \sin \omega t$, a signal $x(t)$ of arbitrary form whose magnitude is much smaller than E_0, the amplitude of the persistent oscillation. The essential result can be stated as follows: If we ignore higher-order modulation terms for the reason that they are negligibly small, or that they are to be suppressed ultimately by suitable filtering, the relay with the input

$$E_0 \sin \omega_0 t + x(t)$$

where $x(t)$ is small compared with E_0, behaves, as far as the transmission of the signal $x(t)$ is concerned, substantially as a linear system having a constant frequency response given by Eq. (6.14).

6.5 Oscillating Control Servomechanisms with Built-in Oscillation. We are now ready to discuss the oscillating control servomechanism. We have already seen that the only effect of the persistent oscillation, as far as the transmission of signals is concerned, is to make the relay into an effectively linear element, with a positive real frequency response. Hence we might have considered the relay to be such an element from the beginning, avoided all explicit mention of the oscillation $E_0 \sin \omega_0 t$, and dealt with the system entirely by means of the concepts and methods given in the earlier chapters. This is, in fact, the novel and excellent procedure proposed by J. C. Lozier.

For the sake of simplicity, we have assumed so far that the inherent properties of the servo afford all of the filtering that is necessary to suppress the unwanted modulation terms introduced by the relay. It is at least conceivable that, in practice, it may sometimes be necessary to supply additional filtering by means of supplementary filters. Naturally, whatever effective filters there may be in the system must be such that they pass the wanted signals. This, in combination with our other considerations, implies that the frequency ω_0 must be above the important parts of the Fourier spectra of the signals.

No matter what filtering we may introduce into the system, the output will always contain, as one component, an oscillation having the frequency ω_0. It is worth remarking that it may not even be desirable to reduce the amplitude of this component below a certain level, by

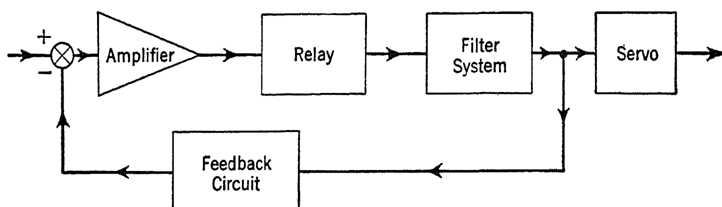


FIG. 6.3

filtering. In fact, such an oscillation furnishes "dynamical lubrication," which diminishes the effects of static friction, backlash, and other parasitic nonlinearities tending to degrade the performance of servomechanisms.

We have not said anything very specific about the way in which the oscillation $E_0 \sin \omega_0 t$ is supplied to the relay; we have merely remarked incidentally that it may be supplied by a subsidiary oscillator. Systems in which the oscillation is supplied in that way have certain advantages in the way of flexibility. However, they have the disadvantage of involving a certain amount of extra equipment. We shall now give a brief description of a variety of oscillating control servomechanisms in which the servomechanism itself is made to supply the oscillation.

Consider the system sketched in Fig. 6.3. Suppose that the system is so designed that in the absence of the input signal it oscillates at a frequency ω_0 determined by the phase shifts of the linear elements in the feedback loop. As we have seen, the relay behaves, as regards the oscillation, as an effectively linear element, having a frequency response which decreases as the amplitude is increased. The amplitude of the oscillation adjusts itself so that the amplification around the loop, determined by the amplifications through the relay and through the linear elements, is unity.

Now suppose that the system is subjected to an input signal. If the corresponding error signal at the input of the relay is sufficiently small, the amplification of the relay for the persistent oscillation is substantially unaffected, and the system continues to oscillate at substantially the original frequency and amplitude. As we have seen, the relay behaves, as regards the signals, as an effectively linear element, having an amplification which is 6 db less than the amplification for the persistent oscillation. It is clear that under these conditions we have an oscillating-control mechanism such as we have discussed above. The only novelty in the present situation is the fact that the frequency and amplitude of the persistent oscillation $E_0 \sin \omega_0 t$ are determined by the system itself, instead of being determined independently, as we have tacitly assumed heretofore.

In all our considerations of the system as a servomechanism, we need only ascribe the proper effective frequency response to the relay, and then proceed in the ways described in the preceding chapters. We do not need to take account explicitly of the persistent oscillation. However, the requirement that the system shall also function as an oscillator imposes certain restrictions on what we can do toward improving its performance as a servomechanism. This can be seen as follows.

Let $F(s)$ denote the transfer function of the feedback loop for the signals computed by using Eq. (6.14). Then the transfer function for the persistent oscillation is $2F(s)$; and, by the very fact that the system does oscillate, there is a purely imaginary root $s = i\omega_0$ for the system transfer function $1 + [1/2F(s)]$. Therefore

$$2F(i\omega_0) = -1$$

Hence, the curve $1/F(s)$ in the Nyquist diagram, as s traces the imaginary axis, is constrained to pass through the point -2. On the other hand, in order that the performance of the system as a servomechanism shall be satisfactory, the curve must meet the conditions we have discussed in Chap. 4, including the condition of avoiding the neighborhood of the point -1. Obviously, the constraint to which the curve is subjected makes it more difficult to meet these conditions than it is in the other systems, where no such constraint exists. In this sense, these self-oscillating servomechanisms are less flexible than are oscillating control servomechanisms in which the oscillation is supplied by an independent generator.

An elementary precaution to be observed, in order that the curve, which is constrained to pass through the point -2, shall avoid the neighborhood of the point -1, is that the curve should intersect the real axis at the point -2 perpendicularly. This implies that the vector

$1/F(i\omega)$ should be varying slowly in magnitude and rapidly in angle at the frequency at which the system oscillates.

6.6 General Oscillating Control Servomechanism. A relay is a nonlinear device. But by mixing the signal with a sinusoidal oscillation of high frequency and large amplitude, the output is made to be linear with respect to the signal. Thus the essential concept of oscillating control servomechanisms is the linearization of a nonlinear system. J. M. Loeb[1] has shown that this concept is applicable to any nonlinear system, and he calls this method the general linearizing process for nonlinear control systems. We shall call the resulting servomechanism the *general oscillating control servomechanism.*

Let us consider a general function $y(x)$, where y is the output and x is the input. If for the variable x we substitute the sum $x + \epsilon$, where ϵ is much smaller than x, then, if the function $y(x)$ is regular, we can expand $y(x + \epsilon)$ into a Taylor series as

$$y(x + \epsilon) = y(x) + \epsilon \left(\frac{dy}{dx}\right)_x + \epsilon^2 \frac{1}{2}\left(\frac{d^2y}{dx^2}\right)_x + \cdots \tag{6.15}$$

We now specify the input x as a periodic function of time t with the period T, and ϵ as a constant. Then it is clear that $y(x)$ is also a periodic function of time with the same period T. The same is true for dy/dx and d^2y/dx^2. Periodic functions can be expanded into Fourier series; thus if we neglect powers of ϵ higher than the first, we have

$$y(x + \epsilon) \approx a_{00} + \sum_{n=1}^{\infty} (a_{0n} \cos n\omega t + b_{0n} \sin n\omega t) + \epsilon\left[a_{10} + \sum_{n=1}^{\infty} (a_{1n} \cos n\omega t + b_{1n} \sin n\omega t)\right] \tag{6.16}$$

where $\omega = 2\pi/T$, the frequency of the input x.

If ϵ is not exactly a constant but a slowly varying function of t such that its fundamental frequency is very much lower than ω, then Eq. (6.16) is still approximately correct. Now consider $y(x)$ as the output-input relation of the nonlinear device, $\epsilon(t)$ as the signal, and $x(t)$ as the superimposed high-frequency, large-amplitude oscillation, not necessarily sinusoidal. The signal information in the output of the nonlinear device is represented by the second term of Eq. (6.16). Since the frequency ω is much higher than that of $\epsilon(t)$, the periodic function represented by the Fourier series

$$a_{10} + \sum_{n=1}^{\infty} (a_{1n} \cos n\omega t + b_{1n} \sin n\omega t)$$

[1] J. M. Loeb, *Ann. de Télécommunications*, **5**, 65–71 (1950).

can be considered as the carrier, while $\epsilon(t)$ is the modulating function. While the above discussion is based upon a direct relation $y(x)$ between the input and output of the nonlinear device, Loeb has shown that Eq. (6.16) is also true for the more general *functional* relation between y and x, that is, y at t is dependent not only upon the instantaneous x at the same t, but also upon all past values of x. This extended concept of input-output relation includes hysteresis effects such as gear blacklash, and is sufficient for almost all nonlinear devices in practice. Therefore, for a general oscillating-control servomechanism, the signal in the output of the nonlinear device appears as a modulated carrier wave. Moreover, the input-output relation is linear as far as the signal is concerned.

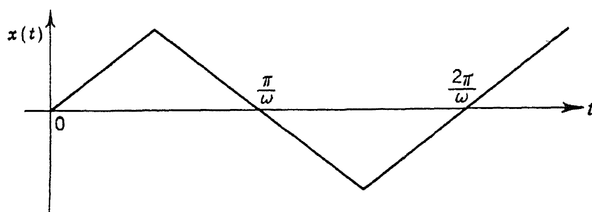


FIG. 6.4

Now let us assume that the superimposed oscillation has a symmetrical wave form such as the sine wave or the saw-tooth wave shown in Fig. 6.4. Then if $y(x)$ is even, or

$$\left.\begin{aligned} y(x) &= y(-x) \\ \text{and} \qquad \left(\frac{dy}{dx}\right)_x &= -\left(\frac{dy}{dx}\right)_{-x} \end{aligned}\right\} \tag{6.17}$$

we have the following relations for the periodic function $y(x)$ and dy/dx

$$\left.\begin{aligned} y(x)_t &= y(x)_{t+\frac{T}{2}} \\ \text{and} \qquad \left(\frac{dy}{dx}\right)_t &= -\left(\frac{dy}{dx}\right)_{t+\frac{T}{2}} \end{aligned}\right\} \tag{6.18}$$

These requirements then specify

$$a_{01} = b_{01} = 0, \qquad a_{10} = 0 \qquad \text{for } y \text{ even} \tag{6.19}$$

Therefore, if the higher harmonics are neglected, the carrier is a sinusoidal oscillation of frequency ω. This is the case of the a-c servomechanism discussed in the previous sections. The design method described there can then be applied to this class of general oscillating control servomechanisms. If $y(x)$ is odd, or $y(x) = -y(-x)$, then a set of conditions similar to Eqs. (6.17) and (6.18) can be written down, and

$$a_{00} = 0, \; a_{11} = b_{11} = 0 \qquad \text{for } y \text{ odd} \tag{6.20}$$

By neglecting higher harmonics, the case is thus identical to the oscillating control servomechanism discussed in Sec. 6.4.

The preceding discussion shows that the characteristics of a nonlinear component in a servomechanism can be linearized by the technique of adding a persistent oscillation to the input and thus converting the system into an oscillating control servomechanism of improved performance. Furthermore, such servomechanisms can be designed with methods already explained in this chapter.

CHAPTER 7

SAMPLING SERVOMECHANISMS

All the servomechanisms that we have considered so far are designed to deal with signals which are given as functions of the continuous variable t. There are situations, however, in which the signals which a servomechanism has to deal with are given as functions of a discrete variable. Such a situation arises, for example, when we have an input signal which has been obtained by determining the values of a function $x(t)$ at equally spaced instants $0, t_0, 2t_0, \ldots$. In such a case, the input signal is not defined at all in the open intervals between the successive sampling instants.

Naturally, when we have a situation of the kind just described, we are interested in the values of the output signal at the sampling instants. Consequently, the servomechanism should function so that the corrective effect which it applies to the output signal is governed only by those values and not by the values which the output signal may have during the intervening intervals. A servomechanism which is designed to function in this way may be called a sampling servomechanism. In this chapter, we give a brief account of a theory of linear sampling servomechanisms which is very similar in its point of view and procedure to the theory of continuously operating servomechanisms that we have been discussing in the preceding chapters.

7.1 Output of a Sampling Circuit. The prototype sampling servomechanism which we shall consider is shown in Fig. 7.1. The system contains the usual forward and feedback circuits. The essential novelty of the system lies in the fact that the feedback path contains a switch, which is operated periodically so that the feedback loop is closed only during short time intervals located at the equally spaced instants $0, t_0, 2t_0, \ldots$. The location of the energy-storing, or frequency-selective, elements in the system affects the theory in matters of detail only. Hence we take the opportunity to simplify the exposition somewhat by assuming that the transfer function of the forward circuit is independent of the frequency. Then it is essential that the switch be placed in the position shown.

The following analysis is based upon the assumption that the intervals during which the switch is closed are so short that the feedback circuit

can be considered to be subjected to a sequence of impulses. It is also based upon the assumption that the response $h_2(t)$ of the feedback circuit to an impulse is a continuous function of time. This means that for small values of t, $h_2(t)$ behaves like t^n, where $n \geq 1$. Then $h_2(t)$ has no jump in value at $t = 0$. If $F_2(s)$ is the transfer function of the feedback circuit, then according to the general formula, Eq. (2.18),

$$h_2(t) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} F_2(s) e^{ts}\, ds \tag{7.1}$$

where γ is a real constant which is greater than the real part of any pole of $F_2(s)$. If for large values of s, $F_2(s)$ behaves like $1/s^m$, then according

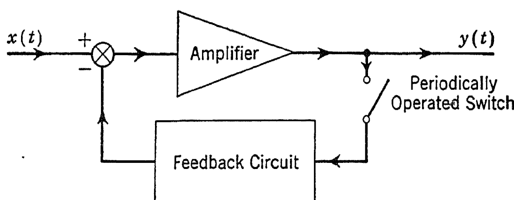


Fig. 7.1

to the "dictionary" of Laplace transforms, Table 2.1, $h_2(t)$ will behave like t^{m-1} for small t. Our condition for the continuity of $h_2(t)$ at $t = 0$ thus requires m to be at least 2. Thus for large s, $F_2(s)$ approaches zero at least as rapidly as $1/s^2$.

Now it is clear that if the input signal $x(t)$ vanishes identically for negative values of t, the value of the output signal $y(t)$ at the typical sampling instant nt_0 is computed as the sum of the effects of all previous impulses and is given by the formula

$$y(nt_0) = F_1 \left[x(nt_0) - \theta t_0 \sum_{k=0}^{n} y(kt_0) h_2(nt_0 - kt_0) \right] \tag{7.2}$$

where F_1 denotes the transfer function of the forward circuit, a constant, and θ is the fraction of the switching cycle during which the switch is closed. $\theta t_0 y(kt_0)$ is thus the "impulse" input to the feedback circuit at $t = kt_0$.

When $x(t)$ and $h_2(t)$ are known at the sampling instants, the values of $y(0)$, $y(t_0)$, $y(2t_0)$, . . . can be calculated successively by Eq. (7.2) in an elementary way. However, instead of proceeding in that way, we shall follow a more illuminating course, which will bring the theory of the sampling servomechanism into a form similar to that of the theory of the ordinary servomechanism discussed in Chap. 4. This approach is due to G. R. Stibitz and C. E. Shannon.

7.2 Stibitz-Shannon Theory. Let us write

$$X^*(s) = \sum_{n=0}^{\infty} x(nt_0)e^{-nt_0s} \tag{7.3}$$

$$Y^*(s) = \sum_{n=0}^{\infty} y(nt_0)e^{-nt_0s} \tag{7.4}$$

and

$$F_2^*(s) = t_0 \sum_{n=0}^{\infty} h_2(nt_0)e^{-nt_0s} \tag{7.5}$$

These functions are thus periodic functions of s, with the imaginary period $2\pi i/t_0$. The functions of nt_0 are thus the Fourier coefficients. The step of going from the functions $x(t)$, $y(t)$, and $h_2(t)$ to $X^*(s)$, $Y^*(s)$, and $F_2^*(s)$ is very similar to the formation of the corresponding Laplace transforms $X(s)$, $Y(s)$, and $F_2(s)$, as indicated by Eq. (2.1). Here the continuous time variable is replaced by the discrete time instants nt_0, and thus the integral sign is replaced by the summation sign. Therefore Eqs. (7.3) to (7.5) represent the natural adaptation of the Laplace-transform technique to the problem of the sampling servomechanism.

For the time being we shall confine our attention to the case in which all of the poles of $F_2(s)$ lie to the left of the imaginary axis. Then the function $h_2(t)$ ultimately decays exponentially as t tends toward infinity, and the series in Eq. (7.5) converges for all values of s with real parts which are greater than a certain negative constant. Of course, the convergence of the series in Eqs. (7.3) and (7.4) depends upon the nature of the input signal. We restrict our attention to input signals for which the series converge in the same manner as the series in Eq. (7.5). This amounts only to the mild sort of restriction on $x(t)$ that we are accustomed to assume in transient theory.

Multiplying Eq. (7.2) through by e^{-nt_0s} and then summing over all values of n, we obtain

$$Y^*(s) = F_1\left[X^*(s) - \theta t_0 \sum_{n=0}^{\infty} e^{-nt_0s} \sum_{k=0}^{n} y(kt_0)h_2(nt_0 - kt_0)\right]$$

But

$$\begin{aligned}\theta t_0 \sum_{n=0}^{\infty} e^{-nt_0s} \sum_{k=0}^{n} y(kt_0)h_2(nt_0 - kt_0) \\ &= \theta t_0 \sum_{n=0}^{\infty} \sum_{k=0}^{n} y(kt_0)e^{-kt_0s}e^{-(n-k)t_0s}h_2(nt_0 - kt_0) \\ &= \theta t_0 \sum_{m=0}^{\infty} \sum_{k=0}^{\infty} y(kt_0)e^{-kt_0s}e^{-mt_0s}h_2(mt_0) \\ &= \theta F_2^*(s)Y^*(s)\end{aligned}$$

The step of changing the summation over n and k to a summation over k and $m = n - k$ is indicated in Fig. 7.2, where the shaded region is the region of summation. Therefore

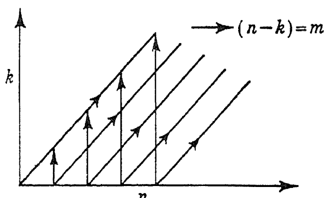


Fig. 7.2

$$Y^*(s) = F_1[X^*(s) - \theta F_2^*(s)Y^*(s)]$$

or

$$\frac{Y^*(s)}{X^*(s)} = \frac{F_1}{1 + \theta F_1 F_2^*(s)} \tag{7.6}$$

Equation (7.6) is the analogue of the basic equation (4.3) for feedback servomechanisms, discussed before. What difference there is between the cases lies in the analytical natures of the functions involved. We shall discuss this point later.

Now let us assume that the character of $y(nt_0)$ is such that the series of $Y^*(s)$ is convergent for values of s with nonnegative real parts. Then the series is convergent for purely imaginary s. Let $s = i\omega$. Then

$$Y^*(i\omega)e^{int_0\omega} = \sum_{m=0}^{\infty} y(mt_0)e^{-it_0\omega(m-n)}$$

Therefore

$$\int_{\omega_0}^{\omega_0+(2\pi/t_0)} Y^*(i\omega)e^{int_0\omega}\, d\omega = y(nt_0)\frac{2\pi}{t_0}$$

or

$$y(nt_0) = \frac{t_0}{2\pi i}\int_{\omega_0}^{\omega_0+(2\pi/t_0)} Y^*(i\omega)e^{int_0\omega} i\, d\omega$$

By putting $i\omega = s$ and $i\omega_0 = s_0$, we have, finally,

$$y(nt_0) = \frac{t_0}{2\pi i}\int_{s_0}^{s_0+(2\pi i/t_0)} Y^*(s)e^{nt_0 s}\, ds$$

By the Cauchy theorem of complex integration,

$$y(nt_0) = \frac{t_0}{2\pi i}\int_\Gamma Y^*(s)e^{nt_0 s}\, ds = \frac{t_0}{2\pi i}\int_\Gamma \frac{F_1 X^*(s)e^{nt_0 s}}{1 + \theta F_1 F_2^*(s)}\, ds \tag{7.7}$$

where, as shown in Fig. 7.3, Γ is a path of integration joining two points separated by the distance $2\pi/t_0$ on the imaginary axis in the s plane, and passing to the right of all singular points of the integrand. This description of Γ is now in such a general form that if $Y^*(s)$ has poles with positive real parts, Eq. (7.7) is still true.

Because of the periodicity of $X^*(s)$ and $F_2^*(s)$, we can add to Γ the dotted lines parallel to the real axis and leave the value of the integral unchanged. The combined path then encloses all the poles of the

integrand. But it can be shown that for reasonable input $X^*(s)$ has no poles with positive real parts. Then the only possible source of unstable output is a zero of the denominator of the integrand of Eq. (7.7) with a positive real part. Thus, in a way very similar to the requirement on conventional servomechanisms, the necessary and sufficient condition for stability is that the equation

$$1 + \theta F_1 F_2^*(s) = 0 \qquad (7.8)$$

shall have no roots in the right half of the s plane. We shall now show how we can implement this condition by an appropriate adaptation of the Nyquist criterion of Sec. 4.3.

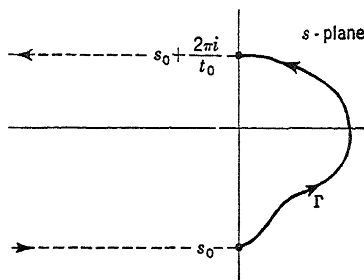


FIG. 7.3

7.3 Nyquist Criterion for Sampling Servomechanisms. Because of the periodicity of $F_2^*(s)$, it suffices to determine whether or not Eq. (7.8) has any roots in a horizontal half strip, of width $2\pi/t_0$, extending to the right from the imaginary axis. We are assuming that $F_2^*(s)$ has no singular points on or to the right of the imaginary axis, and we also assume that $1 + \theta F_1 F_2^*(s)$ has no zero on the imaginary axis. We can, and do, assume that the half strip is adjusted vertically so that $1 + \theta F_1 F_2^*(s)$ has no zeros on the horizontal sides. Now let the point s describe the

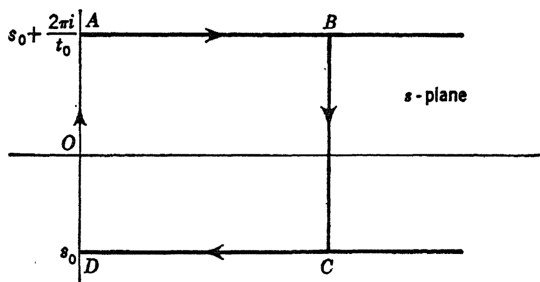


FIG. 7.4

closed curve $ABCDA$ of Fig. 7.4. Then the corresponding tip of the vector $\theta F_1 F_2^*(s)$ describes a certain closed curve, such as $A'B'C'D'A'$ shown in Fig. 7.5. We do not try to show the curve described by the vector realistically. When s describes AB, we have the arc $A'B'$. When s describes BC, we have an arc $B'C'$; and, because of the periodicity of $F_2^*(s)$, $B'C'$ is a closed curve. When s describes CD, we have an arc $C'D'$; and, because of the periodicity of $F_2^*(s)$, $C'D'$ coincides, except for sense, with $A'B'$. Finally, when s describes DA, we have the arc $D'A'$, which is a closed curve.

By Cauchy's theorem, Eq. (7.8) does, or does not, have roots in the rectangle $ABCD$ according to whether the radius vector from the point -1 to the running point $\theta F_1 F_2^*(s)$ does, or does not, make a nonzero net number of revolutions as the running point describes the curve $A'B'C'D'A'$. Now consider what happens when the side BC of the rectangle in Fig. 7.4 recedes to infinity. It is easy to see from Eq. (7.5) that the closed curve formed by the arc $B'C'$ in Fig. 7.5 shrinks to a single point. Therefore the effective curve is the arc $D'A'$. Obviously, Eq. (7.8) does, or does not, have roots in the half strip according to whether the radius vector from the point -1 to the curve does, or does not, make a nonzero net number of revolutions as the limiting curve is described. This is the Nyquist criterion for simple sampling servomechanisms.

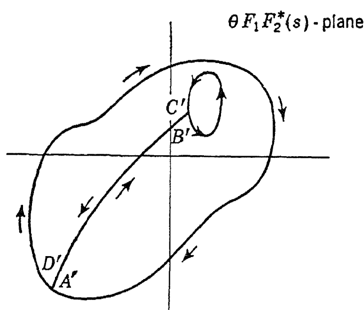


FIG. 7.5

We have now given the fundamentals of the Stibitz-Shannon theory of sampling servomechanisms, which is very similar, both in its point of view and in its form, to the theory of servomechanisms with continuous operation.

7.4 Steady-state Error. If the input is a unit step function,

$$X^*(s) = \sum_{n=0}^{\infty} e^{-nt_0 s} = \sum_{n=0}^{\infty} (e^{-t_0 s})^n = \frac{1}{1 - e^{-t_0 s}}$$

then, according to Eq. (7.7),

$$y(nt_0) = \frac{t_0 F_1}{2\pi i} \int_\Gamma \frac{e^{nt_0 s}\, ds}{[1 - e^{-t_0 s}][1 + \theta F_1 F_2^*(s)]}$$

As $n \to \infty$, the only pole of importance is at the origin,

$$\lim_{n\to\infty} y(nt_0) = \frac{F_1}{1 + \theta F_1 F_2^*(0)} \tag{7.9}$$

This equation gives the "steady-state" output for a constant input of unit magnitude. Therefore the condition for small steady-state error is

$$F_1 \approx 1 + \theta F_1 F_2^*(0)$$

or

$$F_1 \approx \frac{1}{1 - \theta F_2^*(0)} \tag{7.10}$$

This gives the approximate magnitude of the gain for the forward circuit if the output is to follow the input accurately. Equation (7.10) for sampling servomechanisms is the analogue of Eq. (4.11) for continuous servomechanisms.

7.5 Calculation of $F_2^*(s)$. We have one more step to take before our theory of sampling servomechanisms can be regarded as being of much practical value. Both of the functions $F_2(s)$ and $F_2^*(s)$ serve to characterize the feedback circuit: $F_2(s)$ is the significant characteristic when the circuit is used as part of a continuously operating servomechanism; and $F_2^*(s)$ is the significant characteristic when the circuit is used as part of a sampling servomechanism. The familiar function $F_2(s)$ is mathematically much the simpler of the two, and furthermore it is the function which is used directly in our common techniques for designing circuits. It is important, therefore, that we relate $F_2^*(s)$ to $F_2(s)$, and in as direct a manner as possible. We note again that $F_2^*(s)$ is periodic in s with the imaginary period $i2\pi/t_0$. Furthermore, the analysis of the performance of the system by the Nyquist method requires the "frequency response" $F^*(i\omega)$ only for $-\pi/t_0 < \omega < \pi/t_0$.

Assuming that the real part of s is greater than γ, which according to our assumptions is a negative number, we have, by Eqs. (7.1) and (7.5),

$$\begin{aligned} F_2^*(s) &= \frac{t_0}{2\pi i} \sum_{n=0}^{\infty} e^{-nt_0 s} \int_{\gamma - i\infty}^{\gamma + i\infty} F_2(q) e^{nt_0 q}\, dq \\ &= \frac{t_0}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} F_2(q)\, dq \sum_{n=0}^{\infty} e^{-nt_0(s-q)} = \frac{t_0}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} \frac{F_2(q)\, dq}{1 - e^{-t_0(s-q)}} \end{aligned} \tag{7.11}$$

We proceed to evaluate the right-hand member of (7.11) by the method of residues.

The integrand has certain poles: the poles of $F_2(s)$ lie to the left of the path of integration, and the poles which are the roots of the equation $1 - e^{-t_0(s-q)} = 0$ lie to the right of the path of integration. It is easily seen that the integration upward along the line $\gamma - i\infty$ to $\gamma + i\infty$ is equivalent to integration in a clockwise direction along the closed curve formed by that line and the infinite semicircle in the right-half plane. Hence the right-hand member of Eq. (7.11) is $-t_0$ times the sum of the residues of the integrand with respect to the several roots of the equation $1 - e^{-t_0(s-q)} = 0$.

Now the typical root of the equation is $q = s + (2\pi i m/t_0)$, where m is an integer and the residue of the integrand with respect to that pole is $-(1/t_0)F_2[s + (2\pi i m/t_0)]$. Therefore, finally,

$$F_2^*(s) = \sum_{m=-\infty}^{m=\infty} F_2\left(s + \frac{2\pi i m}{t_0}\right) \tag{7.12}$$

This formula gives considerable insight into the properties of $F_2^*(s)$ and at times may be useful in making approximate calculations. However, we can easily obtain an exact representation of $F_2^*(s)$ in finite form.

The function $F_2(s)$ can be represented as the sum of a finite number of partial fractions, thus:

$$F_2(s) = \sum_{k=1}^{n} \frac{a_k}{s - s_k} \tag{7.13}$$

where the a_k's and the s_k's are constants and n is the degree of the denominator polynomial of $F_2(s)$. Consequently, we can write, by using Eq. (7.12),

$$\begin{aligned} F_2^*(s) &= \sum_{k=1}^{n} a_k \left\{\frac{1}{s - s_k} + \sum_{m=1}^{\infty} \left[\frac{1}{(2\pi i m/t_0) + (s - s_k)} - \frac{1}{(2\pi i m/t_0) - (s - s_k)}\right]\right\} \\ &= \sum_{k=1}^{n} a_k \left[\frac{1}{s - s_k} + \sum_{m=1}^{\infty} \frac{2(s - s_k)}{(4\pi^2 m^2/t_0^2) + (s - s_k)^2}\right] \end{aligned} \tag{7.14}$$

Now it is known that coth z has the following expansion:

$$\coth z = \frac{1}{z} + 2z \sum_{m=1}^{\infty} \frac{1}{m^2\pi^2 + z^2}$$

Therefore the sum over m in Eq. (7.14) can be carried out, and we have

$$F_2^*(s) = \frac{t_0}{2} \sum_{k=1}^{n} a_k \coth\left[\frac{(s - s_k)t_0}{2}\right] \tag{7.15}$$

By means of this formula we can compute $F_2^*(s)$ exactly for any value of s.

When t_0 is very small and $F_2(i\omega)$ is negligibly small outside of the interval $-\pi/t_0 < \omega < \pi/t_0$, the qualitative nature of $F_2^*(i\omega)$ is immediately apparent from Eq. (7.12). In fact, $F_2^*(i\omega)$ is approximately equal, in the interval $-\pi/t_0 < \omega < \pi/t_0$, to the function $F_2(i\omega)$. We shall now see that when t_0 is large we can obtain an equally simple approximation

to $F_2^*(i\omega)$. Let us write the roots s_k as

$$s_k = -\lambda_k + i\omega_k \tag{7.16}$$

where the λ's and ω's are all real. In accordance with our assumption, the λ's are all positive. Now we have, according to Eq. (7.15),

$$\begin{aligned} F_2^*(i\omega) &= \frac{t_0}{2} \sum_{k=1}^{n} a_k \coth \left\{ \frac{t_0}{2} [\lambda_k + i(\omega - \omega_k)] \right\} \\ &= \frac{t_0}{2} \sum_{k=1}^{n} a_k \frac{1 + e^{-t_0[\lambda_k + i(\omega - \omega_k)]}}{1 - e^{-t_0[\lambda_k + i(\omega - \omega_k)]}} \cdot \end{aligned}$$

Therefore, for large values of t_0,

$$F_2^*(i\omega) \approx \frac{t_0}{2} \sum_{k=1}^{n} a_k \{1 + 2e^{-t_0[\lambda_k + i(\omega - \omega_k)]}\} \tag{7.17}$$

When s is large, Eq. (7.13) can be written as

$$F_2(s) = \frac{1}{s} \sum_{k=1}^{n} a_k + \frac{1}{s^2} \sum_{k=1}^{n} a_k s_k + \cdots$$

But we have assumed as a condition for continuous response of the feedback circuit to an impulse that $F_2(s) \sim 1/s^2$ when s is large. Therefore

$$\sum_{k=1}^{n} a_k = 0 \tag{7.18}$$

Then Eq. (7.17) becomes

$$F_2^*(i\omega) \approx t_0 e^{-it_0\omega} \sum_{k=1}^{n} a_k e^{t_0 s_k} \tag{7.19}$$

For physical systems, the s_k's are real or form pairs of complex conjugates. Therefore the finite sum in Eq. (7.19) is actually real, and the graph of $F_2^*(i\omega)$ as ω goes from $-\pi/t_0$ to π/t_0 is a circle with the radius

$$\left| t_0 \sum_{k=1}^{n} a_k e^{t_0 s_k} \right| \tag{7.20}$$

7.6 Comparison of Continuously Operating with Sampling Servomechanisms. For small t_0, we have seen that $F_2^*(i\omega)$ is approximately $F_2(i\omega)$. The stability criterion for the continuously operating servomechanism is that the curve $F_1F_2(i\omega)$ should avoid the point -1. For

the sampling servomechanism, the curve $\theta F_1 F_2^*(i\omega)$ should avoid the point -1, or the curve $F_1 F_2(i\omega)$ should avoid the point $1/\theta$. Therefore, if stability is the only consideration, the sampling servomechanism can have much larger gain than the conventional servomechanism.

For large values of t_0, because of Eqs. (7.19) and (7.20), the Nyquist criterion becomes simply

$$\theta F_1 \left| t_0 \sum_{k=1}^{n} a_k e^{t_0 s_k} \right| < 1$$

Since s_k has a negative real part, the radius of the $F_2^*(i\omega)$ curve is very small. This fact together with the smallness of θ, the fraction of time when the switch is closed, allows very large gain for the forward circuit without instability. Thus for any value of the switching period t_0, the condition for stable operation of a sampling servomechanism is very much less stringent than that for a continuously operating feedback servomechanism. Perhaps this is to be expected, because the time interval when feedback of the output actually occurs is very brief, and no restriction is imposed on the output other than that at the switching instant.

7.7 Pole of $F_2(s)$ at Origin. In practice, the function $F_2(s)$ is quite likely to have a pole at $s = 0$. So far we have excluded this case from consideration, in order to avoid having to deal with certain minor complications. However, we shall now consider the case briefly.

In the first place, we observe that when $s = 0$ is a pole of $F_2(s)$ the constant γ must be positive and that our representation of $F_2^*(s)$ in terms of infinite series is valid only for values of s with positive real parts. In the second place, the Nyquist diagram for the system also undergoes changes. Specifically, instead of getting an actually closed curve, we get an open curve, the ends of which are to be regarded as being joined by an infinite semicircle in the clockwise sense. Our representation of $F_2^*(s)$ in finite form, however, remains valid. If we set $s_1 = 0$, then Eq. (7.15) gives

$$F_2^*(i\omega) = \frac{t_0}{2}\left[-ia_1 \cot\frac{\omega t_0}{2} + \sum_{k=2}^{n} a_k \coth\frac{t_0[\lambda_k + i(\omega - \omega_k)]}{2} \right]$$

For t_0 large, we obtain an expression similar to Eq. (7.17), *i.e.*,

$$F_2^*(i\omega) = \frac{t_0}{2}\left[-ia_1 \cot\frac{\omega t_0}{2} + \sum_{k=2}^{n} a_k \{1 + 2e^{-t_0[\lambda_k + i(\omega - \omega_k)]}\} \right]$$

But according to Eq. (7.18),

$$a_2 + a_3 + \cdots + a_n = -a_1$$

so that

$$F_2^*(i\omega) = -\frac{a_1 t_0}{2}\left[1 + i \cot \frac{\omega t_0}{2}\right] + t_0 e^{-it\omega} \sum_{k=2}^{n} a_k e^{t_0 s_k} \qquad (7.21)$$

The constant a_1 is, of course, real and positive. As ω varies from $-\pi/t_0$ to $+\pi/t_0$, the first term gives a vertical straight line parallel to the imaginary axis. The other part of $F_2^*(i\omega)$ is a sinusoidal function. Hence the Nyquist diagram is a sinuous variation of a straight line.

CHAPTER 8

LINEAR SYSTEMS WITH TIME LAG

In this chapter we shall introduce another new element into our linear systems with constant coefficients: the time lag. By time lag τ we mean that the relation between the different variables of the system cannot be expressed as a relation of these variables all taken at some time instant t; but on the contrary, the relation involves some variables taken at the time instant t, and some taken at an earlier instant $t - \tau$. Those taken at the instant $t - \tau$ then lag by the interval τ behind the variables taken at the instant t. This time lag is thus quite different from the characteristic time constant of the first-order linear system introduced in Sec. 3.1. Time-lag systems are represented by differential difference equations of constant coefficients and are more complex than the linear systems studied previously, which are represented by differential equations. Systems with time lag were studied by many investigators: for instance, A. Callander, D. Hartree, and A. Porter,[1] and N. Minorsky.[2] Our interest here is, however, somewhat more restricted. We wish to know: How can we analyze the performance of a feedback servomechanism if there is a characteristic time lag τ in the system? We wish, specifically, to modify the Nyquist method of Sec. 4.3 to apply to time-lag systems.

We shall develop the theory by treating a particular example of such systems, namely, the example of stabilizing the combustion in a rocket motor by feedback control. The problem of combustion instability in rocket motors has been treated by many authors, but the following analysis of combustion lag time originates from the work of L. Crocco.[3] For simplicity of calculation,[4] we shall consider only the case of so-called low-frequency oscillation in a rocket motor using a single liquid propellant.

8.1 Time Lag in Combustion. Let $\dot{m}_b(t)$ be the mass rate of generation of hot gas by combustion at the time instant t. The mass rate of injection at t can be denoted by $\dot{m}_i(t)$. Let $\tau(t)$ be the time lag for that

[1] A. Callander, D. Hartree, and A. Porter, *Trans. Roy. Soc. London* (*A*), **235**, 415–444 (1935).

[2] N. Minorsky, *J. Appl. Mechanics* (ASME), **9**, 67–71 (1942).

[3] L. Crocco, *J. Am. Rocket Soc.* **21**, 163–178 (1951).

[4] The following discussion is based upon a paper in *J. Am. Rocket Soc.*, **22**, 256–262 (1952).

parcel of propellant which is burned at the instant t. Then the mass burned during the interval from t to $t + dt$ must be equal to the mass injected during the time $t - \tau$ to $t - \tau + d(t - \tau)$. Therefore

$$\dot{m}_b(t)\, dt = \dot{m}_i(t - \tau) d(t - \tau) \tag{8.1}$$

The mass of hot gas generated is either used to fill the combustion chamber by raising its pressure $p(t)$ or is discharged through the rocket nozzle. If the frequency of the possible oscillations within the chamber is low, then the pressure in the chamber can be considered uniform, and, as a first approximation,[1] the flow through the nozzle can be considered quasi-stationary. Therefore the mass rate of discharge through the nozzle is proportional to the density of hot gas in the rocket motor. But for a "monopropellant" rocket motor, the temperature of the hot gas is nearly independent of the combustion pressure, and the density of the hot gas is proportional to pressure only. Thus if $\bar{\dot{m}}$ is the steady mass rate of flow through the system, $\bar{M}_g$ is the average mass of hot gas in the motor, $\bar{p}$ is the steady-state pressure in the combustion chamber, and if the volume occupied by the unburned liquid propellant is neglected, we have

$$\dot{m}_b\, dt = \bar{\dot{m}} \left(\frac{p}{\bar{p}}\right) dt + d\left(\bar{M}_g \frac{p}{\bar{p}}\right) \tag{8.2}$$

We now introduce the nondimensional variables φ and μ for the chamber pressure and the rate of injection, respectively, defined as

$$\varphi = \frac{p - \bar{p}}{\bar{p}} \qquad \mu = \frac{\dot{m}_i - \bar{\dot{m}}}{\bar{\dot{m}}} \tag{8.3}$$

φ and μ are then the fractional deviations of the pressure and injection rates from the average. With Eq. (8.3), $\dot{m}_b$ can be eliminated from Eqs. (8.1) and (8.2), and

$$\frac{\bar{M}_g}{\bar{\dot{m}}} \frac{d\varphi}{dt} + \varphi + 1 = \left(1 - \frac{d\tau}{dt}\right) [\mu(t - \tau) + 1] \tag{8.4}$$

To calculate the quantity $d\tau/dt$, Crocco's concept of the pressure dependence of time lag has to be introduced. If the rate at which the liquid propellant is prepared for the final rapid transformation into hot gas is a function $f(p)$, then the lag τ is determined by

$$\int_{t-\tau}^{t} f(p)\, dt = \text{const.} \tag{8.5}$$

This constant can be thought of as the amount of heat that has to be added to a unit mass of the cold injected propellant before "ignition"

[1] H. S. Tsien, *J. Am. Rocket Soc.*, **22**, 139–143 (1952).

occurs. Then $f(p)$ has the physical meaning of the rate of heat transfer from the hot combustion gas to the injected liquid propellant. By differentiating Eq. (8.5) with respect to t,

$$[f(p)]_t - [f(p)]_{t-\tau}\left(1 - \frac{d\tau}{dt}\right) = 0$$

We can now introduce explicitly the concept of a small perturbation from the uniform steady state. Assume that the deviation of the pressure p from the steady-state value $\bar{p}$ is small. Then $f(p)$ at the instant t and $f(p)$ at the instant $t - \tau$ can be expanded as Taylor's series around $\bar{p}$. By taking only the first-order terms,

$$[f(p)]_t = f(\bar{p}) + \bar{p}\left(\frac{df}{dp}\right)_{p=\bar{p}} \varphi(t)$$

$$[f(p)]_{t-\tau} = f(\bar{p}) + \bar{p}\left(\frac{df}{dp}\right)_{p=\bar{p}} \varphi(t - \tau)$$

Here τ is the lag at the average pressure $\bar{p}$, a constant now. Then

$$1 - \frac{d\tau}{dt} = 1 + \left(\frac{d \log f}{d \log p}\right)_{p=\bar{p}} [\varphi(t) - \varphi(t - \tau)] \tag{8.6}$$

By combining Eqs. (8.4) and (8.6), the following equation is obtained:

$$\frac{d\varphi}{dz} + \varphi = \mu(z - \delta) + n[\varphi(z) - \varphi(z - \delta)] \tag{8.7}$$

where

$$n = \left(\frac{d \log f}{d \log p}\right)_{p=\bar{p}} \tag{8.8}$$

and

$$\theta_g = \frac{\bar{M}_g}{\bar{\dot{m}}} \qquad z = \frac{t}{\theta_g} \qquad \delta = \frac{\tau}{\theta_g} \tag{8.9}$$

θ_g is thus the average gaseous mass in the motor divided by the average rate of mass flow through the motor, and is thus the average time between the instant of production by combustion of the hot gas to the instant of discharge through the nozzle. θ_g is therefore called the gas transit time. We shall measure time by this fundamental time constant in the following calculations. z is the nondimensional time variable, and δ is the nondimensional constant time lag of combustion.

If n is a constant independent of $\bar{p}$, then $f(p)$ is proportional to p^n. This is the form of $f(p)$ assumed by Crocco. The present formulation of the problem is slightly more general, in that $f(p)$ is arbitrary and the value of n is to be computed by using Eq. (8.8) and is a function of $\bar{p}$. If $f(p)$ is considered as the rate of heat transfer from the hot combustion

gas to the propellant droplet, then the physical laws of heat transfer indicate a value for n between $\frac{1}{2}$ and 1.

8.2 Satche Diagram. Crocco called the instability of combustion with a constant rate of injection the intrinsic instability. If the injection rate is a constant not influenced by the chamber pressure p, then $\mu \equiv 0$. Therefore the stability problem is controlled by the following simple equation, obtained from Eq. (8.7):

$$\frac{d\varphi}{dz} + (1 - n)\varphi(z) + n\varphi(z - \delta) = 0 \tag{8.10}$$

Equation (8.10) could be treated by the Laplace-transform technique in the same way as equations without time lag discussed in the previous chapters. In fact this is the method used by H. I. Ansoff.[1] However, for the present problem of stability of combustion, the fundamental equation has no forcing term. Thus a more direct approach is that of the classical method of solving linear differential difference equations. That is, let

$$\varphi(z) \sim e^{sz}$$

then

$$s + (1 - n) + ne^{-\delta s} = 0 \tag{8.11}$$

This is the equation for the exponent s. Stability of combustion then requires that the real part of s be negative.

Equation (8.11) can also be obtained by applying the Laplace-transform method to Eq. (8.10). By multiplying Eq. (8.10) by e^{-sz} and then integrating with respect to z from $z = 0$ to $z = \infty$, we have, noting that $\Phi(s)$ is the Laplace transform of $\varphi(z)$,

$$s\Phi(s) - \varphi(0) + (1 - n)\Phi(s) + n \int_0^\infty \varphi(z - \delta)e^{-sz}\, dz = 0$$

But

$$\begin{aligned}\int_0^\infty \varphi(z - \delta)e^{-sz}\, dz &= e^{-s\delta} \int_0^\infty \varphi(z - \delta)e^{-s(z-\delta)}\, dz \\ &= e^{-s\delta}\left[\Phi(s) + \int_{-\delta}^{0} \varphi(z')e^{-sz'}\, dz'\right]\end{aligned}$$

Therefore, if the initial conditions are such that $\varphi = 0$ for $z \leq 0$ for the so-called null initial conditions, then

$$[s + (1 - n) + ne^{-s\delta}]\Phi(s) = 0$$

Hence we have Eq. (8.11). s has the same "meaning" as the variable s in the previous chapters. The only difference between the two is the fact that here s is made nondimensional by the transit time θ_g. It is also interesting to note that if Eq. (8.10) were a nonhomogeneous equa-

[1] H. I. Ansoff, *J. Appl. Mechanics* (ASME), **16**, 158–164 (1949).

tion with a forcing term at the right side, then after applying the Laplace transform to the equation, the resultant equation would also be nonhomogeneous. Then it will be seen that the function

$$F(s) = \frac{1}{s + (1 - n) + ne^{-s\delta}}$$

is the transfer function of the system, with $\varphi(z)$ considered as an output. $F(s)$ is thus another example of a transcendental transfer function.

Crocco determined the value of the complex root s by solving the set of two equations for the real and the imaginary parts of Eq. (8.11).

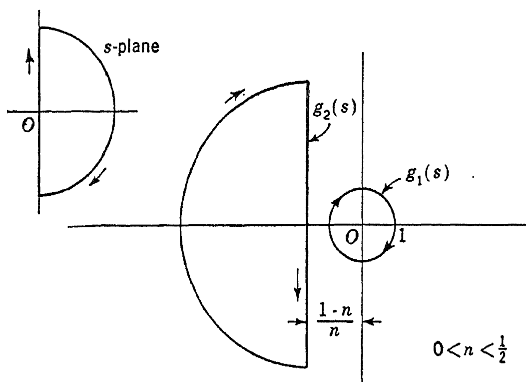


Fig. 8.1

However, if the point of interest is whether the system is stable or not, one can again use Cauchy's theorem of Sec. 4.3 with advantage. Let

$$G(s) = e^{-\delta s} - \left(-\frac{1-n}{n} - \frac{s}{n}\right) \tag{8.12}$$

Then the question of stability is determined by whether $G(s)$ has zeros in the right half of the complex s plane. This question itself can be answered in turn by watching the argument of $G(s)$ when s traces a curve enclosing the right-half s plane, as shown in Fig. 4.4. If the vector $G(s)$ makes a number of complete clockwise revolutions, then, according to Cauchy's theorem, that number is the difference between the number of zeros and the number of poles of $G(s)$ in the right-half s plane. Since $G(s)$ evidently has no poles in the s plane, the number of revolutions of $G(s)$ is the number of zeros. Hence, for stability, the vector $G(s)$ must not make any complete revolutions as s traces the specified curve. Therefore the stability question can be answered by plotting the Nyquist diagram.

A direct application of this method to $G(s)$ as given by Eq. (8.12), however, is inconvenient because of the complication caused by lag

term $e^{-\delta s}$. M. Satche[1] proposed a very elegant and ingenious method to treat such a system with time lag: Instead of treating $G(s)$, break it into two parts,

$$G(s) = g_1(s) - g_2(s) \tag{8.13}$$

where

$$g_1(s) = e^{-\delta s}$$
$$g_2(s) = -\frac{1-n}{n} - \frac{s}{n} \tag{8.14}$$

The vector $G(s)$ is thus a vector with its vertex in $g_1(s)$ and its tail on $g_2(s)$. The graph of $g_1(s)$ for s on the imaginary axis is the unit circle. For s on the large half circle, $g_1(s)$ is within the unit circle. The graph of $g_2(s)$ is the straight line (Fig. 8.1) parallel to the imaginary axis when s is on the imaginary axis. When s is on the large half circle, $g_2(s)$ is a half of a great circle closing the curve on the left. A moment's reflection will show that in order for the vector $G(s)$ not to make complete revolutions for any value of the time lag δ, the $g_2(s)$ curve must lie completely outside the $g_1(s)$ curve. That is, for unconditional intrinsic stability,

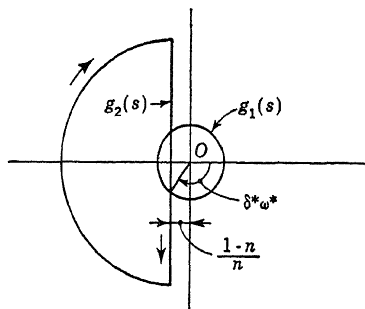


FIG. 8.2

$$\frac{1-n}{n} > 1 \quad \text{or} \quad \tfrac{1}{2} > n > 0 \tag{8.15}$$

It is easily seen now that the separation of $G(s)$ into two parts $g_1(s)$ and $g_2(s)$ allows a great simplification of the respective curves. The diagram of the combination $g_1(s)$ and $g_2(s)$ will be called the *Satche diagram.*

When $n > \frac{1}{2}$, the $g_1(s)$ and $g_2(s)$ curves intersect. Stability is still possible, however, if for $g_2(s)$ within the unit circle of Fig. 8.2 $g_1(s)$ is to the right of $g_2(s)$. This condition is satisfied if

$$\cos\left(\delta\sqrt{2n-1}\right) > -\frac{1-n}{n}$$

or if

$$\delta < \delta^*$$

where

$$\delta^* = \frac{\cos^{-1}\left(-\dfrac{1-n}{n}\right)}{\sqrt{2n-1}} = \frac{1}{\sqrt{2n-1}}\left[\pi - \cos^{-1}\left(\frac{1-n}{n}\right)\right] \tag{8.16}$$

[1] M. Satche, *J. Appl. Mechanics* (ASME), **16,** 419–420 (1949).

When $\delta = \delta^*$, then with

$$\omega^* = \sqrt{2n - 1} \tag{8.17}$$

$G(i\omega^*) = 0$. Therefore, when $\delta = \delta^*$, φ has an oscillatory solution with the frequency ω^*. δ^* and ω^* are thus the nondimensional critical time lag and the nondimensional critical frequency, respectively.

8.3 System Dynamics of a Rocket Motor with Feedback Servo. Consider now a system including the propellant feed and a feedback servo, represented by Fig. 8.3. In order to approximate the elasticity of the feed line, a spring-load capacitance is put at the point midway between the propellant pump and the injector. Near the injector there

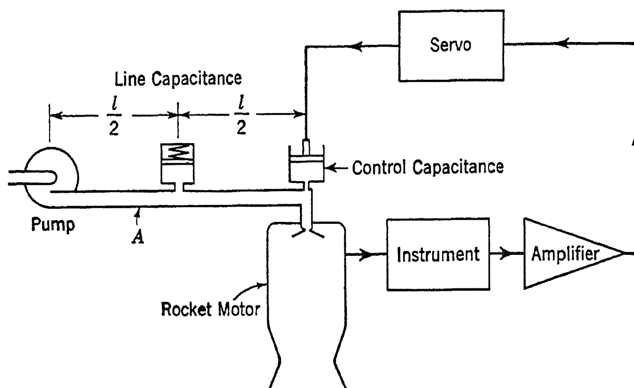


Fig. 8.3

is another capacitance controlled by the servo. The servo receives its signal from the chamber-pressure pickup through an amplifier. If the feed system and the motor design are fixed by the designer, the question is whether it is possible to design an appropriate amplifier so that the whole system will be stable. Because there is no accurate information on the time lag of combustion, a practical design should specify unconditional stability, *i.e.*, stability for any value of time lag δ.

Let $\dot{m}_0$ be the instantaneous mass flow rate out of the propellant pump and p_0 be the instantaneous pressure at the outlet of the pump. The average flow rate must be $\bar{\dot{m}}$. The average pressure is $\bar{p}_0$. The pump characteristics can be represented by the following equation:

$$\frac{p_0 - \bar{p}_0}{\bar{p}_0} = -\alpha \frac{\dot{m}_0 - \bar{\dot{m}}}{\bar{\dot{m}}} \tag{8.18}$$

If the time rate of change of mass flow is small in comparison with speed of propagation of elastic waves in the liquid, but large in comparison with the slow rate of change of the rotating speed of the pump, α is simply the slope of the head-volume curve of the pump at *constant* speed

near the steady-state operating point. For conventional centrifugal pumps, α is approximately 1. For displacement pumps, α is very large. For the constant-pressure pump, or the simple pressure feed, α is zero.

Let $\dot{m}_1$ be the instantaneous mass rate of flow after the spring-loaded capacitance, χ be the spring constant of the capacitance, and p_1 the instantaneous pressure at the capacitance. Then

$$\dot{m}_0 - \dot{m}_1 = \rho\chi \frac{dp_1}{dt} \tag{8.19}$$

where ρ is the density of the propellant, a constant.

In the following calculation, the pressure drop in the line caused by frictional forces will be neglected. Then the pressure difference $p_0 - p_1$ is due to the acceleration of the flow only. That is,

$$p_0 - p_1 = \frac{l}{2A} \frac{d\dot{m}_0}{dt} \tag{8.20}$$

where A, a constant, is the cross-sectional area of the feed line, and l is the total length of the feed line. Similarly, if p_2 is the instantaneous pressure at the control capacitance,

$$p_1 - p_2 = \frac{l}{2A} \frac{d\dot{m}_1}{dt} \tag{8.21}$$

If the mass capacity of the control capacitance is C, then

$$\dot{m}_1 - \dot{m}_i = \frac{dC}{dt} \tag{8.22}$$

Since the control capacitance is very close to the injector, the inertia of the mass of propellant between the control capacitance and the injector is negligible. Then

$$p_2 - p = \frac{1}{2} \frac{\dot{m}_i^2}{\rho A_i^2} \tag{8.23}$$

where A_i is the effective orifice area of the injector. A_i can be eliminated from the calculation by noting that at steady state the difference of pressures $\bar{p}_0$ and $\bar{p}$, or $\Delta\bar{p}$, is

$$\bar{p}_0 - \bar{p} = \Delta\bar{p} = \frac{1}{2} \frac{\bar{\dot{m}}^2}{\rho A_i^2} \tag{8.24}$$

Equations (8.18) to (8.24) describe the dynamics of the feed system. By a straightforward process of elimination of variables, a relation between $\dot{m}_i$, p, and C is obtained. To express this relation in nondimensional form, the following quantities are introduced:

$$P = \frac{\bar{p}}{2\Delta\bar{p}} \qquad E = \frac{2\Delta\bar{p}}{\bar{\dot{m}}\theta_g}\rho\chi \qquad J = \frac{l\bar{\dot{m}}}{2\Delta\bar{p}A\theta_g} \tag{8.25}$$

and

$$\kappa = \frac{C}{\bar{m}\theta_g} \tag{8.26}$$

where θ_g, the gas transit time, is given by Eq. (8.9). Then the nondimensional equation relating φ, μ, and κ is

$$\begin{aligned} P\left\{1 + \alpha E\left(P + \tfrac{1}{2}\right)\frac{d}{dz} + \tfrac{1}{2}JE\frac{d^2}{dz^2}\right\}\varphi + \Big\{[1 + \alpha(P + \tfrac{1}{2})] \\ + [\alpha E(P + \tfrac{1}{2}) + J]\frac{d}{dz} + [\tfrac{1}{2}\alpha JE(P + \tfrac{1}{2}) + \tfrac{1}{2}JE]\frac{d^2}{dz^2} + \tfrac{1}{4}J^2E\frac{d^3}{dz^3}\Big\}\mu \\ + \left\{\alpha(P + \tfrac{1}{2})\frac{d}{dz} + J\frac{d^2}{dz^2} + \tfrac{1}{2}\alpha JE(P + \tfrac{1}{2})\frac{d^3}{dz^3} + \tfrac{1}{4}J^2E\frac{d^4}{dz^4}\right\}\kappa = 0 \end{aligned} \tag{8.27}$$

where z is the nondimensional time variable defined by Eq. (8.9).

The dynamics of the servo control is specified by the composite of the instrument characteristics of the pressure pickup, the response of the amplifier, and the properties of the servo. Since we do not propose to discuss the detailed design of the feedback servo, the over-all dynamics of the servo control is represented by the following operator equation:

$$F\left(\frac{d}{dz}\right)\varphi = \kappa \tag{8.28}$$

where F is the ratio of two polynomials with the denominator of higher order than the numerator.

Equations (8.7), (8.27), and (8.28) are the three equations for the three variables φ, μ, and κ. Since they are equations with constant coefficients, the appropriate forms for the variables are

$$\varphi = ae^{sz} \qquad \mu = be^{sz} \qquad \kappa = ce^{sz} \tag{8.29}$$

By substituting Eq. (8.29) into Eqs. (8.7), (8.27), and (8.28), three homogeneous equations for a, b, and c are obtained. Thus we have

$$\begin{gathered} a[s + (1 - n) + ne^{-\delta s}] - be^{-\delta s} = 0 \\ P\left\{1 + \alpha E(P + \tfrac{1}{2})s + \tfrac{1}{2}JEs^2\right\}a + \Big\{[1 + \alpha(P + \tfrac{1}{2})] \\ + [\alpha E(P + \tfrac{1}{2}) + J]s + [\tfrac{1}{2}\alpha JE(P + \tfrac{1}{2}) + \tfrac{1}{2}JE]s^2 + \tfrac{1}{4}J^2Es^3\Big\}b \\ + s\left\{\alpha(P + \tfrac{1}{2}) + Js + \tfrac{1}{2}\alpha JE(P + \tfrac{1}{2})s^2 + \tfrac{1}{4}J^2Es^3\right\}c = 0 \\ F(s)a - c = 0 \end{gathered}$$

In order for a, b, and c to be nonzero, the determinant formed by their coefficients must vanish. This condition can be written as follows:

$$
\begin{aligned}
&[s + (1 - n)] \left\{ \tfrac{1}{4} J^2 E s^3 + \tfrac{1}{2} JE [1 + \alpha (P + \tfrac{1}{2})] s^2 \right. \\
&\quad \left. + [\alpha E (P + \tfrac{1}{2}) + J] s + [1 + \alpha (P + \tfrac{1}{2})] \right\} \\
&\quad + e^{-\delta s} \left[\tfrac{1}{4} n J^2 E s^3 + \left\{ \tfrac{1}{2} nJE [1 + \alpha (P + \tfrac{1}{2})] + \tfrac{1}{2} JEP \right\} s^2 \right. \\
&\quad + \left\{ n [\alpha E (P + \tfrac{1}{2}) + J] + \alpha EP (P + \tfrac{1}{2}) \right\} s \\
&\quad + \left\{ n [n + \alpha (P + \tfrac{1}{2})] + P \right\} \\
&\quad \left. + sF(s) \left\{ \tfrac{1}{4} J^2 E s^3 + \tfrac{1}{2} \alpha JE (P + \tfrac{1}{2}) s^2 + Js + \alpha (P + \tfrac{1}{2}) \right\} \right] = 0
\end{aligned}
\tag{8.30}
$$

This is the equation for determining the exponent s. $F(s)$ is now recognized as the over-all transfer function of the feedback link. The complete system stability depends upon whether Eq. (8.30) gives roots that have positive real parts.

8.4 Instability without Feedback Servo. The system characteristics without the feedback servo can be simply obtained from the basic equation (8.30) by setting $F(s) = 0$. Let it be assumed that the polynomial multiplied into $e^{-\delta s}$ has no zero in the positive-half s plane, as is usually the case. Then Eq. (8.30) can be divided by that polynomial without introducing poles in the positive-half s plane into the resultant function. That is, for the Satche diagram, one has again

$$G(s) = g_1(s) - g_2(s) \qquad g_1(s) = e^{-\delta s}$$

$g_1(s)$ is thus again the "unit circle." $g_2(s)$ is now much more complicated:

$$
\begin{aligned}
g_2(s) = &\left(\frac{s}{n} + \frac{1 - n}{n} \right) \left\{ \tfrac{1}{4} J^2 E s^3 + \tfrac{1}{2} JE [1 + \alpha (P + \tfrac{1}{2})] s^2 \right. \\
&\qquad \left. + [\alpha E (P + \tfrac{1}{2}) + J] s + [1 + \alpha (P + \tfrac{1}{2})] \right\} \\
&\div \left[\tfrac{1}{4} J^2 E s^3 - \tfrac{1}{2} JE \left\{ 1 + \alpha (P + \tfrac{1}{2}) + (P/n) \right\} s^2 \right. \\
&\left. + \left\{ \alpha E (P + \tfrac{1}{2}) [1 + (P/n)] + J \right\} s + \left\{ 1 + \alpha (P + \tfrac{1}{2}) + (P/n) \right\} \right]
\end{aligned}
\tag{8.31}
$$

The intercept of $g_2(s)$ when s is purely imaginary is given by setting $s = 0$ in Eq. (8.31), *i.e.*,

$$g_2(0) = -\frac{1 - n}{n} \frac{1 + \alpha(P + \tfrac{1}{2})}{1 + \alpha(P + \tfrac{1}{2}) + (P/n)} \tag{8.32}$$

Since all the parameters n, α, and P are positive, the magnitude of $g_2(0)$ is now smaller than the magnitude of $g_2(0)$ given by Eq. (8.14) for the intrinsic stability problem. Thus the effect of the feed system is to move the $g_2(s)$ curve towards the unit circle of $g_1(s)$ in the Satche diagram. For instance, for $n = \frac{1}{2}$, $g_2(s)$ is just tangent to the unit circle for the intrinsic system without considering the propellant feed. But with the propellant feed system, the $g_2(s)$ curve will intersect the unit circle, and the system will become unstable for time lag δ exceeding a certain finite value. The influence of the feed system is thus always destabilizing. This is further confirmed by considering the asymptote of $g_2(s)$ for large imaginary s, obtained from Eq. (8.31). That is,

$$g_2(i\omega) \approx -\left[\frac{i\omega}{n} + \left(\frac{1-n}{n} - \frac{2P}{Jn^2}\right) + \cdots\right] \qquad \text{for } |\omega| \gg 1 \quad (8.33)$$

Therefore, for large imaginary s, $g_2(s)$ approaches asymptotically a line parallel to the imaginary axis at a distance

$$\frac{1-n}{n} - \frac{2P}{Jn^2}$$

to the left of the imaginary axis. The effect of feed system is again to move $g_2(s)$ towards the unit circle.

It is thus evident that for the parameter n near $\frac{1}{2}$ or larger than $\frac{1}{2}$, it would be impossible to design the system for unconditional stability. In the Satche diagram, the $g_1(s)$ and $g_2(s)$ curves will always intersect without a feedback servo.

8.5 Complete Stability with Feedback Servo. If the polynomial $H(s)$,

$$\begin{aligned} H(s) &= \tfrac{1}{4}J^2Es^3 + \left\{\tfrac{1}{4}JE\,[1 + \alpha\,(P + \tfrac{1}{2})] + (JEP/2n)\right\}s^2 \\ &+ \left\{\alpha E\,(P + \tfrac{1}{2}) + (\alpha EP/n)(P + \tfrac{1}{2})\right\}s + \left\{1 + \alpha\,(P + \tfrac{1}{2}) + (P/n)\right\} \\ &+ [\,sF(s)/n]\left\{\tfrac{1}{4}J^2Es^3 + \tfrac{1}{2}\,\alpha JE\,(P + \tfrac{1}{2})\,s^2 + Js + \alpha\,(P + \tfrac{1}{2})\right\} \qquad (8.34) \end{aligned}$$

which multiplies into $e^{-\delta s}$ in Eq. (8.30) has no poles or zeros in the right-half s plane, then the presence of zeros of the expression in Eq. (8.30) in the right-half s plane can be determined from the Satche diagram with

$$g_1(s) = e^{-\delta s}$$

and

$$\begin{aligned} g_2(s) = -\left(\frac{s}{n} + \frac{1-n}{n}\right)&\Big\{\tfrac{1}{4}J^2E\,s^3 + \tfrac{1}{2}JE\,[1 + \alpha\,(P + \tfrac{1}{2})]\,s^2 \\ &+ [\alpha E\,(P + \tfrac{1}{2}) + J]\,s + [1 + \alpha\,(P + \tfrac{1}{2})]\Big\} \div H(s) \qquad (8.35) \end{aligned}$$

As s traces the path of Fig. 4.4, $g_1(s)$ is again a unit circle. Therefore, if simultaneously the $g_2(s)$ curve is completely outside the unit circle, there can be no root of Eq. (8.30) in the right-half s plane. In other words, if the transfer function $F(s)$ of the servo-control link is so designed as to place the $g_2(s)$ curve completely outside the unit circle (Fig. 8.4), then the system is stabilized for all time lags.

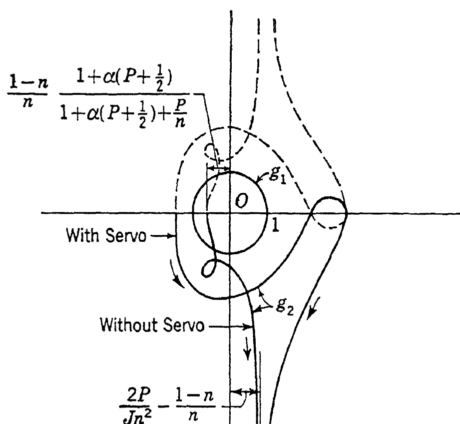


FIG. 8.4

As an example, take

$$n = \tfrac{1}{2} \qquad P = \tfrac{3}{2} \qquad J = 4 \qquad E = \tfrac{1}{4} \qquad \alpha = 1$$

This value of α corresponds to a centrifugal pump for the propellant. Then without the servo control, $g_2(s)$ is

$$g_2(s) = -\frac{1}{2}\frac{(2s+1)(2s^3+3s^2+9s+6)}{s^3+3s^2+6s+6}$$

Of primary interest is the behavior of $g_2(s)$ when s is a purely imaginary number $i\omega$, for ω real. Thus

$$g_2(i\omega) = -\frac{1}{2}\frac{(6-21\omega^2+4\omega^4)(6-3\omega^2)+\omega^2(21-8\omega^2)(6-\omega^2)}{(6-3\omega^2)^2+\omega^2(6-\omega^2)^2} - \frac{1}{2}i\omega\frac{(21-8\omega^2)(6-3\omega^2)-(6-21\omega^2+4\omega^4)(6-\omega^2)}{(6-3\omega^2)^2+\omega^2(6-\omega^2)^2}$$

This curve for $\omega > 0$ is plotted in Fig. 8.5. It is evident that for sufficiently large values of time lag, the system will be unstable. On the other hand, if the $g_2(s)$ curve can be changed by the servo control to, say,

$$g_2(s) = -2\frac{(s+2)(s+3)}{(s+6)}$$

then, as plotted in Fig. 8.5, the new g_2 curve is completely outside the unit circle of $g_1(s)$. Therefore the system is now unconditionally stable. A straightforward calculation from Eqs. (8.31) and (8.35) shows that the required transfer function $F(s)$ for the servo link is

$$F(s) = -4.875 \frac{(s + 1.0528)(s^2 + 0.7164s + 2.6304)}{s(s + 2)(s + 3)(s + 0.5332)(s^2 + 0.4668s + 3.7511)}$$

The servo link has thus the character of an integrating circuit, first discussed in Sec. 3.3. If, with given response of the chamber-pressure pickup and of the servo for the control capacitance, an amplifier could be designed to give an over-all transfer function close to that specified above, the combustion could be stabilized by such a servo control.

As the second example, take

$$n = \tfrac{1}{2} \qquad P = \tfrac{3}{2} \qquad J = 4 \qquad E = \tfrac{1}{4} \qquad \alpha = 0$$

Since $\alpha = 0$, the feed pressure p_0 is thus constant and even when the flow of propellant varies. The case then corresponds to that of a simple pressure feed. Without the feedback servo,

$$g_2(s) = -\frac{1}{2} \frac{(2s + 1)(2s^3 + s^2 + 8s + 2)}{s^3 + 2s^2 + 4s + 4}$$

When s is purely imaginary,

$$g_2(i\omega) = -\frac{1}{2} \frac{(4 - 2\omega^2)(2 - 17\omega^2 + 4\omega^4) + \omega^2(4 - \omega^2)(12 - 4\omega^2)}{(4 - 2\omega^2)^2 + \omega^2(4 - \omega^2)^2} - \frac{1}{2} i\omega \frac{(4 - 2\omega^2)(12 - 4\omega^2) - (4 - \omega^2)(2 - 17\omega^2 + 4\omega^4)}{(4 - 2\omega^2)^2 + \omega^2(4 - \omega^2)^2}$$

This curve of g_2 is plotted in Fig. 8.6. It is evident that without servo-control the combustion will be unstable for sufficiently long time lag. In fact, the system is even less stable than the system considered in the first example: it will become unstable at shorter time lag. The part of the g_2 curve near $\omega = 2$ is of special interest. Near $\omega = 2$, the curve comes so close to the unit circle of g_1 that if the value of time lag δ is such as to make g_1 and g_2 for $\omega \sim 2$ very close to each other, then an almost undamped oscillation at $\omega \sim 2$ can occur. This critical value of δ is evidently smaller than the critical δ determined from the true intersection of g_2 with the unit circle at $\omega \sim 0.65$.

For unconditional stability, g_2 should be displaced out of the unit circle, to, say, the same "stable" curve as in the first example. The required transfer function $F(s)$ is calculated to be

$$F(s) = -4.875 \frac{(s + 0.8126)(s^2 - 0.04337s + 2.6506)}{s^2(s + 2)(s + 3)(s^2 + 4)}$$

The required servo link must then have the character of a double integrating circuit. Furthermore, the transfer function has two purely imaginary poles at $\mp 2i$. This unrealistic requirement on the amplifier comes from the original feed-system dynamics and is due to the neglection of frictional damping in the feed line. In any actual system, the frictional damping in the feed line will remove these purely imaginary

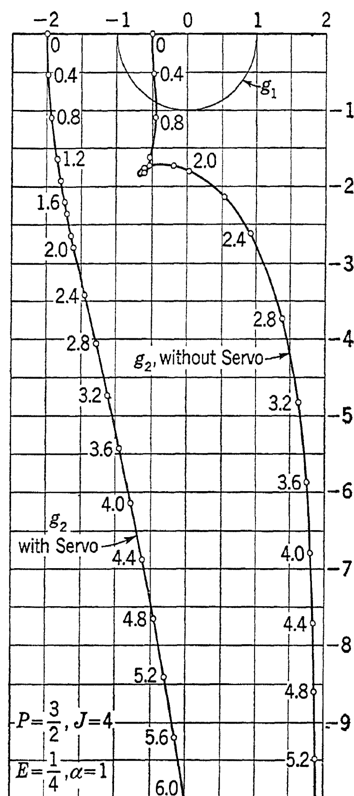


Fig. 8.5

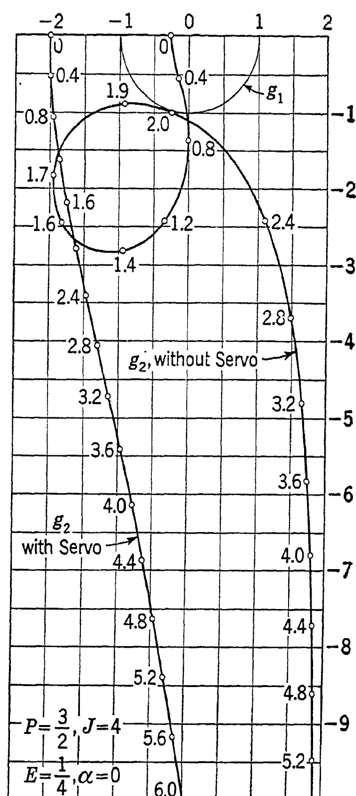


Fig. 8.6

poles of the required transfer function $F(s)$ and replace them by two complex conjugate poles.

It should be emphasized that the advantage of using a feedback servo to stabilize the combustion lies in its great flexibility in being able to obtain unconditional stability with any value of time lag δ or τ. In as much as there are no accurate data on the time lag, the possibility of unconditional stability is of very real, practical importance. Even more than this, servo stabilization also makes it possible to design the system for stability against any expected variation in the parameter n. From physical reasoning, n probably takes a value between $\frac{1}{2}$ and 1. Take

the worst possibility, and design the system for unconditional stability with $n \approx 1$. Then the system will be stable for all expected values of n. Therefore stabilization by feedback servo can be assured without having to know the exact values of the parameters of the system.

8.6 General Stability Criteria for Time-lag Systems. In the preceding discussion of servo stabilization, it is assumed that the polynomial $H(s)$, of Eq. (8.34), has no pole or zero in the right-half s plane. This, however, is not necessarily the case. In general then, one should first investigate the number of zeros and poles of $H(s)$ in the right-half s plane. To do this, it should be recognized that the polynomial in Eq. (8.34) before the factor $F(s)$ usually does not have zeros in the right-half s plane. Therefore, instead of studying $H(s)$, one can study the ratio of $H(s)$ to that polynomial. That is, the number of zeros and poles of $H(s)$ in the right-half s plane is the same as the number of zeros and poles of the following function:

$$H(s) \div \left[\tfrac{1}{4}J^2Es^3 + \left\{ \tfrac{1}{2}JE[1 + \alpha(P + \tfrac{1}{2})] + (JEP/2n) \right\} s^2 + \left\{ \alpha E(P + \tfrac{1}{2}) + (\alpha EP/n)(P + \tfrac{1}{2}) \right\} s + \left\{ 1 + \alpha(P + \tfrac{1}{2}) + (P/n) \right\} \right] = 1 + L(s) \quad (8.36)$$

where

$$L(s) = \frac{1}{n} sF(s) \left[\tfrac{1}{4}J^2Es^3 + \tfrac{1}{2}\alpha JE(P + \tfrac{1}{2})\, s^2 + Js + \alpha\,(P + \tfrac{1}{2}) \right] \div \left[\tfrac{1}{4}J^2Es^3 + \left\{ \tfrac{1}{2}JE[1 + \alpha(P + \tfrac{1}{2})] + (JEP/2n) \right\} s^2 + \left\{ \alpha E(P + \tfrac{1}{2}) + (\alpha EP/n)(P + \tfrac{1}{2}) \right\} s + \left\{ 1 + \alpha(P + \tfrac{1}{2}) + (P/n) \right\} \right] \quad (8.37)$$

According to the Nyquist criterion, the number of poles and zeros for $1 + L(s)$ in the right-half s plane can be found by plotting the Nyquist diagram of $1 + L(s)$ with s tracing the curve of Fig. 4.4. In fact, if $1 + L(s)$ or $H(s)$ has r zeros and q poles in the right-half s plane, then $L(s)$ will carry out $r - q$ clockwise revolutions around the point -1 as s traces the semicircle. Hence the necessary information on $H(s)$ can be obtained by plotting the Nyquist diagram of $L(s)$.

When one divides Eq. (8.30) by $H(s)$ in order to get $g_1(s)$ and $g_2(s)$ as given by Eq. (8.35), q zeros and r poles are introduced in the right-half s plane. The q poles of $L(s)$ must come from $F(s)$, since the polynomial in the denominator of Eq. (8.37) has no zero in the right-half s plane. Therefore the original expression in Eq. (8.30) also has q poles in the right-half s plane. Hence in order for the original expression in

Eq. (8.30) to have no zero in the right-half s plane, $g_2(s)$ must make $-q + (q - r) = -r$ clockwise revolutions around the unit circle. In order for the stability to be unconditional, *i.e.*, stable for all time lag, the $g_2(s)$ curve should never intersect the unit circle. Therefore the *general unconditional stability criteria are, first, the $g_2(s)$ curve must lie completely outside the unit circle; and, second, $g_2(s)$ must make r counterclockwise revolutions around the unit circle as s traces the conventional path enclosing the right-half s plane.* These are the criteria for stability with the Satche

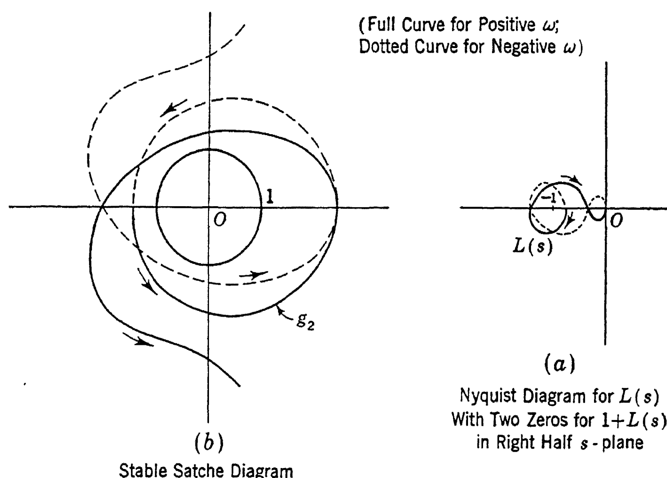


FIG. 8.7

diagram. To determine r, one has to use the Nyquist diagram of $L(s)$, Eq. (8.37). Thus the stability problem for the general case requires both the Satche diagram and the Nyquist diagram (Fig. 8.7).

It is evident that the stability criteria developed here using a combination of the Satche diagram and the Nyquist diagram are applicable to any system with a time lag τ. The stability of such systems always reduces to the question of ascertaining whether there is root of

$$M(s) = 0$$

with a positive real part, where $M(s)$ contains terms with the factor $e^{-\tau s}$. The principle of the method, as seen from the previous discussions, is to divide $M(s)$ by the coefficient of $e^{-\tau s}$ in $M(s)$, such that

$$\frac{M(s)}{1 + L(s)} = G(s) = g_1(s) - g_2(s)$$

and

$$g_1(s) = e^{-\tau s}$$

The curves of $g_1(s)$ and $g_2(s)$ as s traces the right-half semicircle shown in Fig. 4.4 then constitute the Satche diagram, with $g_1(s)$ represented

by the unit circle. To find out whether dividing $M(s)$ by $1 + L(s)$ has introduced roots with positive real parts to the Satche diagram, we have to plot the Nyquist diagram of $L(s)$. The number of roots of $M(s) = 0$ with positive real parts can then be determined by Cauchy's theorem.

The $g_2(s)$ function contains the transfer function of the feedback link with an amplifier which is under the designer's control. $g_2(s)$ may also contain transcendental functions of s coming from other parts of the system. Since the transfer function of the amplifier in the feedback link is generally a ratio of two polynomials in s, it is difficult to compensate completely for the destabilizing effects of a transcendental function. However, in the Satche diagram, the critical part of the $g_2(s)$ path is that part close to the unit circle of $g_1(s)$. But $g_2(s)$ close to the unit circle is generally obtained by small values of s. Hence for the critical part of $g_2(s)$, the transcendental function can be expanded into a power series in s. The amplifier of the feedback link can then be designed using a few terms of the power series as an approximation. Thus the amplifier compensates the destabilizing effects of the system in the critical region. Of course, the performance of the system should be checked finally by the stability criteria developed, using the amplifier design characteristics. This is the procedure suggested by F. E. Marble. For details, the original work[1] should be consulted.

[1] F. E. Marble and D. W. Cox, *J. Am. Rocket Soc.*, **23**, 75–81 (1953).

CHAPTER 9

LINEAR SYSTEMS WITH STATIONARY RANDOM INPUTS

In the previous chapters, the inputs to a system are considered to be definitely specified functions of time t. However, there are many engineering problems for linear systems with constant coefficients where the inputs cannot be so definitely described. An example of such an engineering problem is the problem of the motion and the stresses induced in the structure of an airplane wing in a turbulent air stream. Here the input can be considered to be the time-varying air-flow pattern. The air-flow pattern cannot be described as a definite function of time but has to be recognized as a random function of time, specified by certain statistical characteristics. It is then evident that the output of the system, the stresses in this case, must also be a random function and can also be described only in statistical terms. The first objective of this chapter is then to find a convenient method of calculating the statistical properties of the output from the specified statistical properties of the input. This forms an easy extension of the early investigations by P. Langevin of Brownian motion.

Another example of random input is the so-called *noise* in control signals. The noise is introduced by disturbances and fluctuations beyond the control of the designer. The problem of noise is a subject of much research in connection with communications engineering. There, the central question is how to design the system so that the effects of the unavoidable noise can be minimized and the useful information of the signal not destroyed. We shall discuss this particular problem of noise filtering in Chap. 16. The problem of this chapter is, however, somewhat different. In our present problem, the random output is the only output of the system. Our purpose in the design of the system, particularly the design of the feedback servomechanism, is to obtain with a given input an output of the desired statistical characteristics. We shall see that the transfer-function method developed in the previous chapters remains useful in the present task.

9.1 Statistical Description of a Random Function. Let us consider a system which generates a random function $y_1(t)$. Now to formulate

the concept of a statistical description of such a random function, we have to consider a great number of systems identical to the first. Such a group of systems is called an assembly. The random functions generated by the members of the assembly are $y_1(t)$, $y_2(t)$, $y_3(t)$, ... The random character of the function is exhibited by the fact that, although the systems are identical, the value of the function generated by any member of the assembly at any specified instant of time t is generally different from the value generated by another member at the same time instant. But we can ask for what fraction of the total number of the systems y occurs in a given range y to $y + dy$. This fraction will depend on y and t and will be proportional to dy when dy is small. This fraction is the probability that y will lie between y and $y + dy$ at time t. It is written as $W_1(y,t)\,dy$. The function $W_1(y,t)$ is called the *first probability distribution.* Next we can consider all the pairs of values of y occurring at two given instants t_1 and t_2. The fraction of the total number of pairs in which y occurs in the range y_1 to $y_1 + dy_1$ at t_1 and in the range y_2 to $y_2 + dy_2$ at t_2 is written as $W_2(y_1,t_1;y_2,t_2)\,dy_1\,dy_2$. The function $W_2(y_1,t_1;y_2,t_2)$ is called the *second probability distribution.* Higher probability distributions can be similarly constructed.

The fact that the above formulation of the statistical description of the random function depends upon observations carried out simultaneously on a very large number of identical systems may be objectionable on the ground of practical difficulty in observation. However, if the random function is a *stationary random function* in the sense that all statistical characteristics of the function are time independent, then the large assembly of identical systems is not necessary—all necessary observations can be made on a single system, for a very long period of time. The record of observation can then be cut into pieces of length Θ, Θ being large in comparison with the characteristic time of the function. Then each piece contains the same statistical information about the behavior of the system, since statistically the origin of the time scale is of no consequence. The different pieces can then be considered as an assembly of observations on identical systems, and the various probability distributions can be determined. Furthermore, these distributions now become somewhat simpler: W_1 will be independent of time t, and W_2 will be dependent only upon the time interval $\tau = t_2 - t_1$. Hence for stationary random functions, $W_1(y)\,dy$ is the probability of finding y between y and $y + dy$; $W_2(y_1,y_2;\tau)\,dy_1\,dy_2$ is the probability of finding a pair of values between y_1 and $y_1 + dy_1$ and between y_2 and $y_2 + dy_2$ at an interval of time equal to τ. Since the random functions of engineering problems can very often be considered as stationary random functions, we shall limit the following discussion to such random functions.

It should be emphasized that these probability distributions embody

all the information about the statistical properties of a random function. Or we may say that the probability distributions "define" the random function. Of course, these distributions W_n are not arbitrary but must satisfy the following conditions:

(a) $W_n \geq 0$, because there is no such thing as negative probability;

(b) W_n is symmetric with respect to its variables y_i, *i.e.*,

$$W_2(y_1,y_2;\tau) = W_2(y_2,y_1;\tau) \tag{9.1}$$

This is obvious from the meaning of W_2 as a joint probability distribution.

(c) Higher distributions imply lower distributions. That is,

$$\int_{-\infty}^{\infty} W_2(y_1,y_2;\tau)\, dy_2 = W_1(y_1) = W_1(y) \tag{9.2}$$

where the integration is to be carried over all possible values of y_2. Note that integration over y_2 also eliminates τ. Furthermore,

$$\int_{-\infty}^{\infty} W_1(y)\, dy = 1 \tag{9.3}$$

This simply means that the probability of all occurrences must be a certainty.

9.2 Average Values. From the first probability distribution $W_1(y)$, the *average value* $\bar{y}$ of y can be found:

$$\bar{y} = \int_{-\infty}^{\infty} yW_1(y)\, dy \tag{9.4}$$

Since we have limited ourselves to stationary random functions, the average value can be also obtained by taking the *time average* of $y(t)$. That is,

$$\bar{y} = \lim_{\Theta\to\infty} \frac{1}{\Theta} \int_{-\Theta/2}^{\Theta/2} y(t)\, dt \tag{9.5}$$

The equality of the *assembly average*, Eq. (9.4), and the time average, Eq. (9.5), is a characteristic of stationary random functions. We shall utilize this property repeatedly in the following calculations.

Equation (9.4) can be generalized to arbitrary powers of y. Thus

$$m_n = \overline{y^n} = \int_{-\infty}^{\infty} y^n W_1(y)\, dy \tag{9.6}$$

m_n is called the *nth moment* of the first probability distribution. From the first and second moments, we can compute what is called the *fluctuation, variance*, or *mean deviation* σ:

$$\begin{aligned}\sigma^2 = \overline{(y - \bar{y})^2} &= \int_{-\infty}^{\infty} (y - \bar{y})^2 W_1(y)\, dy \\ &= \int_{-\infty}^{\infty} [y^2 - 2\bar{y}y + (\bar{y})^2] W_1(y)\, dy = \overline{y^2} - (\bar{y})^2\end{aligned} \tag{9.7}$$

σ is thus a measure of the "width" of the probability distribution $W_1(y)$ about the average value $\bar{y}$. Similarly, the third moment gives a measure of the *skewness* of the probability distribution. More and more information about $W_1(y)$ can be deduced as more moments are known. For some cases, the knowledge of moments uniquely determines the distribution. For instance, if

$$\left.\begin{aligned} m_{2k+1} &= 0 \qquad \text{for } k = 0, 1, 2, \ldots \\ m_{2k} &= 1 \cdot 3 \cdot 5 \cdots (2k-1)\sigma^{2k} \end{aligned}\right\} \tag{9.8}$$

then the first probability distribution $W_1(y)$ is the well-known *Gaussian distribution*, or *normal distribution*,

$$W_1(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2/2\sigma^2} \tag{9.9}$$

Sometimes it is convenient to choose the origin of the y coordinate in such a way that $\bar{y}$ vanishes, *i.e.*, the origin is the average value of y. When this is done, we say the probability distribution is *normalized.* In this case, the square of the mean deviation is simply the second moment $\overline{y^2}$, as seen from Eq. (9.7).

The most important average value derived from the second probability distribution $W_2(y_1,y_2;\tau)$ is the *correlation function* $R(\tau)$. It is defined as

$$\begin{aligned} R(\tau) &= \overline{y_1y_2} = \overline{y(t)y(t+\tau)} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1y_2W_2(y_1,y_2;\tau)\, dy_1\, dy_2 \end{aligned} \tag{9.10}$$

For a stationary random function, this is clearly also obtainable from time averaging:

$$R(\tau) = \lim_{\Theta\to\infty} \frac{1}{\Theta} \int_{-\Theta/2}^{\Theta/2} y(t)y(t+\tau)\, dt \tag{9.11}$$

The function $R(\tau)$ thus gives a measure of the interrelation of the y's measured at two different time instants. It is to be expected that as the time interval τ increases, the interrelationship or "memory" must be weakened and, in the end, when τ is very large, $y(t)$ and $y(t+\tau)$ will be independent of each other. Then according to the rule of probability calculus, the second probability distribution is equal to the product of $W_1(y_1)$ and $W_1(y_2)$. Thus for large τ,

$$R(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1y_2W_1(y_1)W_1(y_2)\, dy_1\, dy_2 = (\bar{y})^2 \tag{9.12}$$

For $\tau = 0$, it is obvious from Eq. (9.11) that

$$R(0) = \overline{y^2} \tag{9.13}$$

Now since $R(\tau)$ can be calculated with an arbitrary shift of the origin of the time scale,

$$R(\tau) = \overline{y(t)y(t+\tau)} = \overline{y(t-\tau)y(t)}$$

If we differentiate the equation with respect to τ and then put $\tau = 0$,

$$R'(0) = \overline{y(t)y'(t)} = -\overline{y(t)y'(t)}$$

Therefore,

$$R'(0) = \overline{y(t)y'(t)} = 0 \tag{9.14}$$

In these equations, a prime indicates differentiation with respect to time. Thus the correlation of a random function with its derivative at the *same* time instant is zero. This means that the slope of the record of y at any y has an equal probability of being negative or positive.

If we differentiate Eq. (9.14) twice with respect to τ and then set $\tau = 0$, we have

$$R''(0) = \overline{y(t)y''(t)} = -\overline{y'^2} \tag{9.15}$$

This equation allows us to calculate the mean square of the derivative of y from the correlation function. Similarly, the mean square of the second derivative of y can be calculated as

$$R''''(0) = \overline{y''^2} \tag{9.16}$$

9.3 Power Spectrum. Of special significance for our applications of the theory of random functions is the notion of the spectrum of a random function. Let us suppose that a function $y(t)$ is observed for a long time Θ. If $y(t)$ is assumed to vanish outside the interval, then $y(t)$ can be developed into a Fourier integral[1]

$$y(t) = \int_{-\infty}^{\infty} A(\omega)e^{i\omega t}\,d\omega \tag{9.17}$$

where $A(\omega)$ is the amplitude of the frequency ω. It can be computed from $y(t)$ by the inversion formula:

$$A(\omega) = \frac{1}{2\pi}\int_{-\Theta/2}^{\Theta/2} y(t)e^{-i\omega t}\,dt \tag{9.18}$$

If we denote $A^*(\omega)$ as the complex conjugate of $A(\omega)$, then, since $y(t)$ is real, Eq. (9.18) indicates that

$$A^*(\omega) = A(-\omega) \tag{9.19}$$

Now we can calculate the average $\overline{y^2}$ in terms of $A(\omega)$ as

$$\overline{y^2} = \lim_{\Theta\to\infty} \frac{1}{\Theta}\int_{-\Theta/2}^{\Theta/2} y^2(t)\,dt$$

$$= \lim_{\Theta\to\infty} \frac{1}{\Theta}\int_{-\Theta/2}^{\Theta/2} dt \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\omega\,d\omega'\,A(\omega)A(\omega')e^{i(\omega+\omega')t}$$

[1] See for instance Whittaker and Watson, "Modern Analysis," Sec. 9.7, p. 188, Cambridge-Macmillan, 1943.

By the substitution $\omega'' = -\omega'$, we have

$$\overline{y^2} = \lim_{\Theta \to \infty} \frac{1}{\Theta} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d\omega\, d\omega''\, A(\omega) A^*(\omega'') \int_{-\Theta/2}^{\Theta/2} e^{i(\omega - \omega'')t}\, dt$$

$$= \lim_{\Theta \to \infty} \frac{2}{\Theta} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(\omega) A^*(\omega'') \frac{\sin [\frac{1}{2}(\omega - \omega'')\Theta]}{\omega - \omega''}\, d\omega\, d\omega''$$

If we now introduce the new variable ξ, defined as

$$\xi = \frac{\Theta}{2}(\omega - \omega'')$$

then

$$\omega'' = \omega - \frac{2\xi}{\Theta}$$

Consequently,

$$\overline{y^2} = \lim_{\Theta \to \infty} \frac{2}{\Theta} \int_{-\infty}^{\infty} A(\omega)\, d\omega \int_{-\infty}^{\infty} A^*\left(\omega - \frac{2\xi}{\Theta}\right) \frac{\sin \xi}{\xi}\, d\xi$$

$$= \left[\lim_{\Theta \to \infty} \frac{1}{\Theta} \int_0^{\infty} |A(\omega)|^2\, d\omega\right] 4 \int_{-\infty}^{\infty} \frac{\sin \xi}{\xi}\, d\xi$$

$$= 4\pi \lim_{\Theta \to \infty} \frac{1}{\Theta} \int_0^{\infty} |A(\omega)|^2\, d\omega$$

Therefore, if we put

$$\Phi(\omega) = \lim_{\Theta \to \infty} \frac{4\pi}{\Theta} |A(\omega)|^2 \tag{9.20}$$

then

$$\overline{y^2} = \int_0^{\infty} \Phi(\omega)\, d\omega \tag{9.21}$$

The function $\Phi(\omega)$ is thus a real function and is called the *power spectrum* of the random function. Equations (9.20) and (9.21) enable us to compute the average value $\overline{y^2}$ from the Fourier coefficient $A(\omega)$. This relation is the Parseval theorem.

Let us consider next the correlation function $R(\tau)$. By combining Eqs. (9.11) and (9.17), we have

$$R(\tau) = \lim_{\Theta \to \infty} \frac{1}{\Theta} \int_{-\Theta/2}^{\Theta/2} y(t) y(t + \tau)\, dt$$

$$= \lim_{\Theta \to \infty} \frac{1}{\Theta} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(\omega) A(\omega') e^{i\omega\tau}\, d\omega\, d\omega' \int_{-\Theta/2}^{\Theta/2} e^{i(\omega + \omega')t}\, dt$$

Then, by an argument similar to the above, we have

$$R(\tau) = \int_0^{\infty} \Phi(\omega) \cos \omega\tau\, d\omega \tag{9.22}$$

By setting $\tau = 0$, Eqs. (9.21) and (9.22) reduce to the relation of Eq. (9.13). By differentiating Eq. (9.22) with respect to τ and then setting $\tau = 0$, Eq. (9.15) is obtained. According to the Fourier inversion theorem,

$$\Phi(\omega) = \frac{2}{\pi} \int_0^\infty R(\tau) \cos \omega\tau \, d\tau \tag{9.23}$$

Equations (9.22) and (9.23) allow the computation of either the correlation function or the power spectrum when one of these is known and are called the Wiener-Khintchine relations.

The power spectrum $\Phi(\omega)$ may contain peaks of the Dirac δ-function type. This certainly is the case when $\bar{y}$ is not zero, or, in the terminology of electrical engineering, when there is a d-c term. Then

$$\Phi(\omega) = 2(\bar{y})^2\delta(\omega) + \Phi_1(\omega) \tag{9.24}$$

where $\delta(x)$ is defined as

$$\left.\begin{aligned} &\delta(-x) = \delta(x) = 0 && \text{for } x \neq 0 \\ &\delta(x) \to \infty && \text{for } x = 0 \\ &\text{such that} \quad \int_{-\infty}^{\infty} \delta(x)\, dx = 1 \quad \text{and} && \int_0^\infty \delta(x)\, dx = \tfrac{1}{2} \end{aligned}\right\} \tag{9.25}$$

For pure "noise," the peak at $\omega = 0$, corresponding to the d-c term, will usually be the only peak, so that $\Phi_1(\omega)$ will be a regular function, representing the really continuous spectrum. But it is also possible to have several sinusoidal oscillations superimposed upon the noise. In that case, the power spectrum will have additional peaks at the discrete frequencies of these oscillations.

9.4 Examples of the Power Spectrum. We shall show a few power spectra computed from correlation functions. If the correlation function is given by a Gaussian curve,

$$R(\tau) = R(0)e^{-\alpha^2\tau^2} \tag{9.26}$$

then, according to Eq. (9.23), the corresponding power spectrum is

$$\left.\begin{aligned} \Phi(\omega) &= \frac{2}{\pi} R(0) \int_0^\infty \cos(\omega\tau)e^{-\alpha^2\tau^2}\, d\tau = \Phi(0)e^{-(\omega^2/4\alpha^2)} \\ \text{where} \quad \Phi(0) &= \frac{R(0)}{\alpha\sqrt{\pi}} \end{aligned}\right\} \tag{9.27}$$

It is interesting to note that as $\alpha \to \infty$ the correlation function becomes zero for all finite τ, and $R(0) \to \infty$ in such a manner that $R(\tau)$ becomes a δ function. This means that subsequent y's are not correlated at all and that the random function is the "most chaotic" of all. As $\alpha \to \infty$, the power spectrum is a constant, independent of the frequency. This

most chaotic random function is called the *white noise* and often describes the naturally generated random variations.

Another example is the small isotropic turbulence in a fluid flow of otherwise uniform velocity. It has been shown by von Kármán and Howarth[1] that the fundamental second-order correlation functions are $R_1(\tau)$ and $R_2(\tau)$: $R_1(\tau)$ is the correlation function of the fluctuating velocity parallel to the mean flow direction at the same space point, but at time interval τ apart (Fig. 9.1). $R_2(\tau)$ is the correlation function of the fluctuating velocity normal to the mean flow direction. If U is the mean velocity and L is a scale of turbulence, these correlation functions can be approximately expressed as

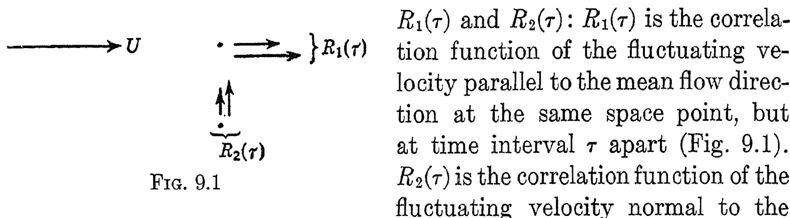


FIG. 9.1

$$R_1(\tau) = R_1(0)e^{-\tau U/L} \tag{9.28}$$

$$R_2(\tau) = R_2(0)e^{-\tau U/L}\left(1 - \frac{1}{2}\frac{\tau U}{L}\right) \tag{9.29}$$

By using Eq. (9.23), the power spectra $\Phi_1(\omega)$ and $\Phi_2(\omega)$ for velocity fluctuations parallel and normal to the mean flow direction, respectively, are

$$\Phi_1(\omega) = \Phi_1(0)\frac{1}{1 + (\omega L/U)^2} \tag{9.30}$$

$$\Phi_2(\omega) = \Phi_2(0)\frac{1 + 3(\omega L/U)^2}{[1 + (\omega L/U)^2]^2} \tag{9.31}$$

where $\Phi_1(0)$ and $\Phi_2(0)$ are the values of the power spectra at $\omega = 0$. They are related to $R_1(0)$ and $R_2(0)$ by

$$\left.\begin{aligned} \Phi_1(0) &= \frac{2}{\pi}\frac{L}{U}R_1(0) \\ \Phi_2(0) &= \frac{1}{\pi}\frac{L}{U}R_2(0) \end{aligned}\right\} \tag{9.32}$$

9.5 Direct Calculation of the Power Spectrum. It is not necessary, of course, to calculate the power spectrum from the correlation function. Sometimes it is possible to determine the spectrum directly from the specified character of the random function $y(t)$ itself. Let us consider, for example, the case where $y(t)$ consists of a series of pulses that have identical shape and a constant repetition frequency but whose heights vary according to some probability distribution. The heights of successive pulses are, however, uncorrelated. This is shown in Fig. 9.2 for a rectangular pulse. If $\eta(t)$ represents a single pulse of unit height, then

$$y(t) = \sum_k a_k \eta(t - kT) \tag{9.33}$$

[1] von Kármán and Howarth, *Proc. Roy. Soc.* (*A*), **164**, 192 (1938).

where T is the spacing of the pulses and a_k the height of the kth pulse. The first step in computing the power spectrum is to determine the Fourier spectrum $A(\omega)$ by Eq. (9.18). Let $\Theta = 2NT$, then

$$A(\omega) = \frac{1}{2\pi}\int_{-NT}^{NT} y(t)e^{-i\omega t}\,dt = \frac{1}{2\pi}\int_{-NT}^{NT}\sum_k a_k\eta(t-kT)e^{-i\omega t}\,dt$$
$$= \sum_{-N}^{N} a_k e^{-i\omega kT}\frac{1}{2\pi}\int_{-\infty}^{\infty}\eta(\xi)e^{-i\omega\xi}\,d\xi = \alpha(\omega)\sum_{-N}^{N} a_k e^{-i\omega kT}$$

where $\alpha(\omega)$ is the Fourier spectrum of the single pulse,

$$\alpha(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\eta(\xi)e^{-i\omega\xi}\,d\xi \tag{9.34}$$

For a rectangular pulse of width 2ϵ and unit height,

$$\alpha(\omega) = \frac{1}{2\pi}\int_{-\epsilon}^{\epsilon} e^{-i\omega\xi}\,d\xi = \frac{1}{\pi}\frac{\sin\omega\epsilon}{\omega} \tag{9.35}$$

According to Eqs. (9.19) and (9.20), the power spectrum is

$$\Phi(\omega) = \frac{4\pi}{T}|\alpha(\omega)|^2 \lim_{N\to\infty}\frac{1}{2N}\left[\sum_{-N}^{N}\sum_{-N}^{N} a_k a_l e^{-i\omega(k-l)T}\right] \tag{9.36}$$

To carry out the limiting process in Eq. (9.36), it will be convenient first to make an assembly average of the whole equation. Since the power

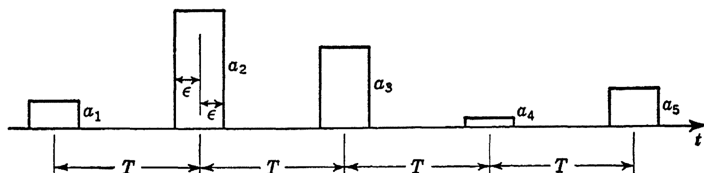


FIG. 9.2

spectrum $\Phi(\omega)$ is the same for every member of the assembly, the left side of Eq. (9.36) is not changed by averaging over the assembly. The right side of Eq. (9.36) will be simplified by such an averaging process. Let $\bar{a}$ be the average of a_k and a_l, and $\overline{a^2}$ the average of the square of a_k and a_l. Then

$$a_k a_l = (a_k - \bar{a})(a_l - \bar{a}) + \bar{a}[(a_k - \bar{a}) + (a_l - \bar{a})] + (\bar{a})^2$$

We substitute this expression into the right side of Eq. (9.36) and then average over the assembly. Then since the successive heights of the pulses are not correlated, the assembly average of $(a_k - \bar{a})(a_l - \bar{a})$ is zero, unless $k = l$. When $k = l$, the assembly average of $(a_k - \bar{a})(a_l - \bar{a})$

is $\overline{a^2} - (\bar{a})^2$. Thus the first term under the limit sign is

$$\lim_{N\to\infty} \frac{1}{2N} \sum_{-N}^{N} \sum_{-N}^{N} \overline{(a_k - \bar{a})(a_l - \bar{a})} e^{-i\omega(k-l)T} = \overline{a^2} - (\bar{a})^2$$

The assembly averages of $(a_k - \bar{a})$ and $(a_l - \bar{a})$ are obviously zero. Therefore, finally,

$$\Phi(\omega) = \frac{4\pi}{T} |\alpha(\omega)|^2 \left\{ [\overline{a^2} - (\bar{a})^2] + (\bar{a})^2 \lim_{N\to\infty} \frac{1}{2N} \left| \sum_{-N}^{N} e^{-i\omega kT} \right|^2 \right\} \tag{9.37}$$

The sum in Eq. (9.37) will be $2N + 1$ for $\omega = 2n\pi/T$ when n is an integer. Hence, in the limit, the second term of Eq. (9.37) will be infinite. For other values of ω the sum will be finite, and for $N \to \infty$ the limiting value will be zero. Clearly then, in the limit, the second term has the character of a series of peaks, or δ functions, at the frequency $2n\pi/T$. To determine the coefficient of these δ functions, we have to calculate the area under the curve for the typical interval $-\pi < \omega T < \pi$. Thus by integrating the sum, the required area is

$$\int_{-\pi/T}^{\pi/T} d\omega \frac{1}{2N} \left| \sum_{-N}^{N} e^{-i\omega kT} \right|^2 = \frac{1}{2N} \int_{-\pi/T}^{\pi/T} \frac{1 - \cos 2N\omega T}{1 - \cos \omega T} d\omega = \frac{2\pi}{T}$$

Since the area under the δ function, according to Eq. (9.25), is unity, the required coefficient is $2\pi/T$. Therefore, finally, the power spectrum for the specified stationary random function is

$$\Phi(\omega) = 2\omega_0 |\alpha(\omega)|^2 \left\{ [\overline{a^2} - (\bar{a})^2] + (\bar{a})^2 \omega_0 \sum_{n=0}^{\infty} \delta(\omega - n\omega_0) \right\} \tag{9.38}$$

where ω_0 is the frequency corresponding to the fundamental period T, *i.e.*,

$$\omega_0 = \frac{2\pi}{T} \tag{9.39}$$

Hence the power spectrum contains a continuous part that has the same shape as the power spectrum of a single pulse. The intensity of this continuous spectrum is determined by the square of the mean deviation σ of the pulse heights. There is, in addition, a discrete spectrum at the frequencies $n\omega_0$, for n an integer, where the intensities are also determined by the spectrum of the single pulse.

Let us consider next a series of pulses that have an identical shape and height but a repetition period varying around an average value T. The spacing between pulses will be $T + \epsilon$, with ϵ distributed according to a

specified probability function $P(\epsilon)$. The average value of ϵ is by implication zero. The successive ϵ's will again be assumed to be uncorrelated. Figure 9.3 shows such a random function with rectangular pulses. Therefore the random function is represented by

$$y(t) = \sum_k \eta(t - kT - \epsilon_k) \tag{9.40}$$

where $\eta(t)$ represents the single pulse. According to Eq. (9.18), with $\Theta = 2NT$,

$$A(\omega) = \alpha(\omega) \sum_{-N}^{N} e^{-i\omega(kT+\epsilon_k)}$$

$\alpha(\omega)$ is the Fourier spectrum of the single pulse given by Eqs. (9.34) and (9.35). The power spectrum of $y(t)$, according to Eqs. (9.19) and

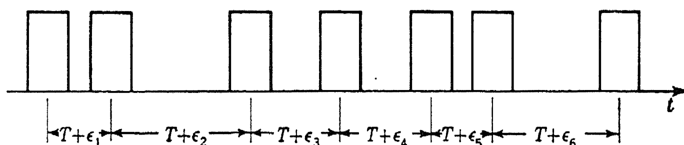


FIG. 9.3

(9.20), is then

$$\Phi(\omega) = \frac{4\pi}{T} |\alpha(\omega)|^2 \lim_{N\to\infty} \frac{1}{2N} \left[\sum_{-N}^{N} \sum_{-N}^{N} e^{-i\omega\epsilon_k} e^{-i\omega\epsilon_l} e^{-i\omega(k-l)T} \right] \tag{9.41}$$

Now let us introduce the function $\chi(\omega)$ defined by

$$\chi(\omega) = \int_{-\infty}^{\infty} P(\epsilon) e^{-i\omega\epsilon}\, d\epsilon \tag{9.42}$$

$\chi(\omega)$ is sometimes called the *characteristic function* of ϵ and is the Fourier transform of $P(\epsilon)$. We shall write then

$$e^{-i\omega\epsilon_k} e^{+i\omega\epsilon_l} = \{[e^{-i\omega\epsilon_k} - \chi(\omega)] + \chi(\omega)\}\{[e^{+i\omega\epsilon_l} - \chi^*(\omega)] + \chi^*(\omega)\}$$

where $\chi^*(\omega)$ is the complex conjugate of $\chi(\omega)$ and is equal to $\chi(-\omega)$. We now substitute this expanded form into Eq. (9.41) and make an assembly average first. The limiting process is greatly simplified by the fact that the ϵ's are not correlated. The final result is

$$\Phi(\omega) = 2\omega_0 |\alpha(\omega)|^2 \left\{ [1 - |\chi(\omega)|^2] + |\chi(\omega)|^2 \omega_0 \sum_{n=0}^{\infty} \delta(\omega - n\omega_0) \right\} \tag{9.43}$$

where ω_0 is the frequency defined by Eq. (9.39). Here the shape of the continuous spectrum and the intensities of the discrete spectrum are no

longer determined solely by the spectrum of the single pulse but depend also on the characteristic function of the distribution of ϵ.

As the third example of direct calculation of the power spectrum, consider the stationary random function $y(t)$ indicated in Fig. 9.4. The function takes the value of either $+1$ or -1 with the interval T. T, however, is not a constant but is distributed according to a specified probability distribution $P(T)$, where $T \geq 0$. It is further specified that the successive time intervals T are not correlated. Let us denote the successive intervals by T_k, where $k = 1, 2, \ldots$, and take the time

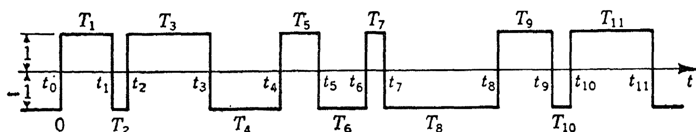


FIG. 9.4

period Θ of integration in Eq. (9.18) to be from $t = 0$ to $t = N\bar{T}$, where $\bar{T}$ is the average time interval defined by

$$\bar{T} = \int_0^\infty TP(T)\, dT \tag{9.44}$$

Then

$$A(\omega) = \frac{1}{2\pi} \int_0^{N\bar{T}} y(t) e^{-i\omega t}\, dt = \frac{1}{2\pi} \frac{1}{i\omega} \sum_{k=1}^{N} (-1)^k (e^{-i\omega t_k} - e^{-i\omega t_{k-1}})$$

where t_k is the time instant at the end of the kth interval. The above expression can be rewritten as

$$A(\omega) = \left[\frac{1}{\pi} \frac{1}{i\omega} \sum_{k=1}^{N} (-1)^k e^{-i\omega t_k} \right] - \frac{1}{2\pi} \frac{1}{i\omega} (-1)^N e^{-i\omega t_N} + 1$$

Hence, by Eq. (9.20), we have the power spectrum as

$$\Phi(\omega) = \frac{4}{\pi \bar{T} \omega^2} \left[\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \sum_{k'=1}^{N} (-1)^{k+k'} e^{-i\omega(t_k - t_{k'})} \right] \tag{9.45}$$

Now let us consider $k > k'$, say $k = k' + m$, then

$$e^{-i\omega(t_k - t_{k'})} = e^{-i\omega T_{k'+1}} e^{-i\omega T_{k'+2}} \cdots e^{-i\omega T_{k'+m}} \tag{9.46}$$

Since the successive intervals are uncorrelated, the probability of occurrence of the product to the right in Eq. (9.46) is the product of the probability of occurrence of each factor. If we introduce the characteristic function $\chi(\omega)$ of the distribution $P(T)$,

$$\chi(\omega) = \phi(\omega) + i\psi(\omega) = \int_0^\infty P(T) e^{-i\omega T}\, dT \tag{9.47}$$

then $\chi(\omega)$ is the average value of $e^{-i\omega T}$. Therefore the assembly average of the product of Eq. (9.46) is simply $[\chi(\omega)]^m$. In the double sum of Eq. (9.45) there are approximately N such products, all with the sign $(-1)^m$. Hence in the limit these products contribute a term $(-1)^m[\chi(\omega)]^m$. m ranges from 1 to ∞ in the limit. The sum of all such terms is then

$$\sum_{m=1}^{\infty} (-1)^m[\chi(\omega)]^m = -\frac{\chi(\omega)}{1 + \chi(\omega)}$$

The contribution from terms with $k' > k$ is just the complex conjugate of the contribution from terms with $k > k'$. These contributions are all there is to the double sum of Eq. (9.45) with one exception, the case $k = k'$. These values of k and k' give a contribution corresponding to 1 within the square bracket. Hence Eq. (9.45) is finally

$$\Phi(\omega) = \frac{4}{\pi\bar{T}\omega^2}\left\{1 - 2\mathcal{R}\left[\frac{\chi(\omega)}{1 + \chi(\omega)}\right]\right\}$$

where $\mathcal{R}$ means the "real part of" the expression. Let the real and the imaginary parts of $\chi(\omega)$ be $\phi(\omega)$ and $\psi(\omega)$, respectively, as shown in Eq. (9.47). Then

$$\Phi(\omega) = \frac{4}{\pi\bar{T}\omega^2}\frac{1 - \phi^2(\omega) - \psi^2(\omega)}{[1 + \phi(\omega)]^2 + \psi^2(\omega)} \tag{9.48}$$

If the distribution $P(T)$ is *Poisson's distribution*,

$$P(T) = \frac{1}{\bar{T}}e^{-T/\bar{T}} \tag{9.49}$$

where $\bar{T}$ is the average time interval defined by Eq. (9.44), then the power spectrum of such a randomly switched function of unit amplitude is

$$\Phi(\omega) = \frac{\bar{T}}{\pi}\frac{1}{1 + (\omega\bar{T}/2)^2} \tag{9.50}$$

The complete lack of any regular periodicity in the random function considered makes the power spectrum continuous and smooth without any of the sharp peaks of the previous examples.

9.6 Probability of Large Deviations from the Mean. If the random function is the stress in a structure, then it is not sufficient to know the average value of the stress. For safety, we shall want to know the probability of occurrence of stresses exceeding the specified working stress of the structural material, *i.e.*, the probability $P[|y| \geq k]$ of the magnitude

of the random function y exceeding the value k. If the first probability distribution function $W_1(y)$ is known, the answer is very simple:

$$P[|y| \geq k] = \int_{-\infty}^{-k} W_1(y)\,dy + \int_{k}^{\infty} W_1(y)\,dy \tag{9.51}$$

But in many engineering problems, the probability distribution function is not known. What is known is the mean value $\bar{y}$ and the mean deviation σ. However, even under these restricted circumstances, it is still possible to give general but broad estimates of the probability of

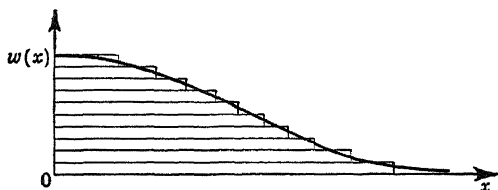


FIG. 9.5

occurrence of large deviations from the mean. For instance, if $g(y)$ is a nonnegative function of y, then since $W_1(y)$ is by definition nonnegative,

$$\overline{g(y)} = \int_{-\infty}^{\infty} g(y)W_1(y)\,dy \geq K \int_{g(y)\geq K} W_1(y)\,dy \tag{9.52}$$

The last integration is to be carried over all y's satisfying the condition $g(y) \geq K$. But the last integral is just $P[g(y) \geq K]$. Thus

$$P[g(y) \geq K] \leq \frac{\overline{g(y)}}{K} \tag{9.53}$$

This is the so-called Chebyshev inequality. Now let

$$g(y) = (y - \bar{y})^2$$

Then according to Eq. (9.7),

$$\overline{g(y)} = \sigma^2 = \overline{y^2} - (\bar{y})^2$$

where σ is the mean deviation from the mean. Let $K = k^2\sigma^2$; then Eq. (9.53) gives the Bienaymé-Chebyshev inequality

$$P[|y - \bar{y}| \geq k\sigma] \leq \frac{1}{k^2} \tag{9.54}$$

The Bienaymé-Chebyshev inequality is known to be too broad for most practical applications, and the upper limit given is, in general, much too high. A sharper estimation can be given for $W_1(y)$ that has only a single maximum, the so-called mode. This estimate for "unimodal" distribution is that due to Gauss. To prove Gauss's inequality, let us consider a function $w(x)$ (Fig. 9.5) which is monotonically

decreasing in the range $x > 0$. It is seen that $w(x)$ can be considered as a superposition of functions which are constant from $x = 0$ to $x = x_0$, and zero for $x > x_0$. Let $v(x) = 1$ for $0 \le x \le x_0$, and $v(x) = 0$ for $x > x_0$. Then for any $K > x_0$,

$$K^2 \int_K^\infty v(x)\,dx = 0$$

But if $0 < K \le x_0$,

$$K^2 \int_K^\infty v(x)\,dx = K^2(x_0 - K)$$

The maximum of this quantity for K within the range specified is $\frac{4}{27}x_0^3$. Thus it is generally true that

$$K^2 \int_K^\infty v(x)\,dx \le \tfrac{4}{9} \int_0^\infty x^2 v(x)\,dx$$

By superposition, we have

$$K^2 \int_K^\infty w(x)\,dx \le \tfrac{4}{9} \int_0^\infty x^2 w(x)\,dx$$

Now consider any unimodal distribution with the abscissa $x = y - y_0$, where y_0 is the mode. Then

$$K^2 \int_K^\infty W_1(x)\,dx \le \tfrac{4}{9} \int_0^\infty x^2 W_1(x)\,dx$$

and

$$K^2 \int_{-\infty}^{-K} W_1(x)\,dx \le \tfrac{4}{9} \int_{-\infty}^0 x^2 W_1(x)\,dx$$

By adding these expressions,

$$K^2 P[|y - y_0| \ge K] \le \tfrac{4}{9}\nu^2$$

where ν is the mean deviation from the mode, defined by

$$\nu^2 = \overline{(y - y_0)^2} = \int_{-\infty}^\infty (y - y_0)^2 W_1(y)\,dy \tag{9.55}$$

Let $K = k\nu$; then we obtain the Gauss inequality

$$P[|y - y_0| \ge k\nu] \le \frac{4}{9k^2} \tag{9.56}$$

If the *distribution is symmetrical*, then $y_0 = \bar{y}$, $\nu = \sigma$, and Eq. (9.56) reduces to

$$P[|y - \bar{y}| \ge k\sigma] \le \frac{4}{9k^2} \tag{9.57}$$

Equation (9.57) is a sharper estimate of the probability than Eq. (9.54).

Often it will be possible to assume, at least approximately, a Gaussian distribution. Then, by using the asymptotic expansion of the error

function, it is easily shown that

$$P[|y - \bar{y}| \geq k\sigma] \approx \frac{e^{-\frac{1}{2}k^2}}{k\sqrt{2\pi}} \qquad \text{for } k \gg 1 \tag{9.58}$$

This is a very small probability. For instance for $k = 3$, the probability is only 0.002. Equation (9.54) will say only that the probability is less than 0.1111, while Eq. (9.57) will say that the probability is less than 0.0493. The difference in these estimates is of course caused by the different degree of information available to the estimation. The more general the assumption, the broader the estimate.

9.7 Frequency of Exceeding a Specified Value. If the random function is the stress in a structure and if the design is to be based upon the repeated occurrences of a certain value of stress, *i.e.*, the "fatigue" stress of the material, then it is necessary to know the probable number of times per unit time the random function will exceed the value $y = \xi$. The quantity is evidently one half of the number of times per unit time the random function will pass through the value ξ. Let $N_0(\xi)$ denote the number of times of passing. This number was first computed by S. O. Rice.[1] We shall follow his method.

Let $W(y,y')\,dy\,dy'$ be the joint probability of having the random function y between y and $y + dy$ and the time derivative y' between y' and $y' + dy'$ at the same time instant. This probability can be also interpreted as the fraction of time per unit time that the random function y and its derivative y' will simultaneously have values within the specified ranges. But crossing the interval dy takes the time $dy/|y'|$. Hence the expected or probable number of crossings at ξ and y' per unit time is equal to the quotient of $W(\xi,y')\,dy\,dy'$ and $dy/|y'|$, or $|y'|W(\xi,y')\,dy'$. The number $N_0(\xi)$ is obtained by integrating over all y'. Thus

$$N_0(\xi) = \int_{-\infty}^{\infty} |y'|W(\xi,y')\,dy' \tag{9.59}$$

But Eq. (9.15) shows that for any differentiable random function, y and y' are not correlated. Then according to the general laws of probability calculation, $W(y,y')$ is simply the product of the first probability distributions $W_1(y)$ and $W(y')$. Hence Eq. (9.59) can be written as

$$N_0(\xi) = W_1(\xi) \int_{-\infty}^{\infty} |y'|W(y')\,dy' \tag{9.60}$$

When $W(y')$ is symmetrical, Eq. (9.60) can be further reduced to

$$N_0(\xi) = 2W_1(\xi) \int_0^{\infty} y'W(y')\,dy' \qquad \text{for } W(y') \text{ symmetrical} \tag{9.61}$$

[1] S. O. Rice, *Bell System Tech. J.*, **23**, 282 (1944); **25**, 46 (1945).

If $W(y')$ is a Gaussian distribution, with the mean deviation σ', then, according to Eq. (9.9),

$$N_0(\xi) = \frac{2W_1(\xi)}{\sigma'\sqrt{2\pi}} \int_0^\infty y' e^{-y'^2/2\sigma'^2}\, dy' = \frac{2\sigma' W(\xi)}{\sqrt{2\pi}} \tag{9.62}$$

The mean deviation σ' can be computed from the power spectrum $\Phi(\omega)$ by using Eqs. (9.15) and (9.22):

$$\sigma'^2 = \int_0^\infty \omega^2 \Phi(\omega)\, d\omega \tag{9.63}$$

If $W_1(y)$ is also a Gaussian distribution with the mean $\bar{y}$ and the mean deviation σ, then with Eqs. (9.7) and (9.21) we obtain

$$N_0(\xi) = \frac{1}{\pi}\frac{\sigma'}{\sigma} e^{-\frac{1}{2}\frac{(\xi-\bar{y})^2}{\sigma^2}} = \frac{1}{\pi} e^{-\frac{1}{2}\frac{(\xi-\bar{y})^2}{\sigma^2}} \left[\frac{\int_0^\infty \omega^2\Phi(\omega)\, d\omega}{\int_0^\infty \Phi(\omega)\, d\omega - (\bar{y})^2}\right]^{\frac{1}{2}} \tag{9.64}$$

This is the formula given by Rice.

9.8 Response of a Linear System to Stationary Random Input. We shall finally give the answer to the question we posed in the introduction to this chapter: Given the stationary random input to a linear system with constant coefficients, what is the output? From the exposition of the elements of the theory of random functions in the previous sections, it is evident that the key to this question is the calculation of the power spectrum of the output from the power spectrum of the input. When the power spectrum of the output is known, we can easily compute the correlation function by Eq. (9.22), and the mean-square value by Eq. (9.21). Then the probability of large deviations from the mean and the frequency of exceeding a specified value can be estimated by methods given in Secs. 9.6 and 9.7. For many engineering problems, knowledge of these characteristics of the output is generally sufficient.

Let the input be $x(t)$ with the power spectrum $\Phi(\omega)$ and the correlation function $R_i(\tau)$. Then by Eqs. (9.21) and (9.22), we have

$$\overline{x^2} = \int_0^\infty \Phi(\omega)\, d\omega = R_i(0) \tag{9.65}$$

and

$$R_i(\tau) = \int_0^\infty \Phi(\omega) \cos \omega\tau\, d\tau = \frac{1}{2}\int_{-\infty}^\infty \Phi(\omega) e^{i\omega\tau}\, d\omega \tag{9.66}$$

where the relation $\Phi(-\omega) = \Phi(\omega)$, as seen from Eq. (9.23), is used. Similarly, let the output be $y(t)$ with the power spectrum $g(\omega)$ and the correlation function $R_0(\tau)$. Then

$$\overline{y^2} = \int_0^\infty g(\omega)\, d\omega = R_0(0) \tag{9.67}$$

and

$$R_0(\tau) = \frac{1}{2} \int_{-\infty}^{\infty} g(\omega) e^{i\omega\tau} \, d\omega \tag{9.68}$$

As before, let $h(t)$ be the response of the linear system to a unit impulse at $t = 0$. For a process started at $t = -\infty$, the output can be written as

$$y(t) = \int_{-\infty}^{t} x(\tau) h(t - \tau) \, d\tau$$

where $x(\tau)\,d\tau$ is the impulse applied at the instant $t = \tau$. Since $h(t) \equiv 0$ for $t < 0$, the upper limit of the integral can be extended to $+\infty$. Then

$$y(t) = \int_{-\infty}^{\infty} x(\tau) h(t - \tau) \, d\tau = \int_{-\infty}^{\infty} x(t - u) h(u) \, du \tag{9.69}$$

The correlation function $R_0(\tau)$ is thus

$$R_0(\tau) = \overline{y(t) y(t + \tau)} = \int_0^{\infty} \int_0^{\infty} \overline{x(t - u) x(t + \tau - u')} h(u) h(u') \, du \, du'$$

But

$$\overline{x(t - u) x(t + \tau - u')} = \overline{x(t) x(t + \tau + u - u')} = R_i(\tau + u - u')$$

Hence by using Eqs. (9.66) and (9.68) we have

$$\int_{-\infty}^{\infty} g(\omega) e^{i\omega\tau} \, d\omega = \int_{-\infty}^{\infty} \int_0^{\infty} \int_0^{\infty} \Phi(\omega) e^{i\omega(\tau + u - u')} h(u) h(u') \, du \, du' \, d\omega \tag{9.70}$$

Now if $F(s)$ is the transfer function of the linear system, it is the Laplace transform of $h(t)$. Thus [*cf.* Eq. (3.50)],

$$F(i\omega) = \int_0^{\infty} e^{i\omega u} h(u) \, du$$

Hence Eq. (9.70) can be now considered as

$$\int_{-\infty}^{\infty} g(\omega) e^{i\omega\tau} \, d\omega = \int_{-\infty}^{\infty} \Phi(\omega) F(i\omega) F(-i\omega) \, e^{i\omega\tau} \, d\omega$$

Therefore the power spectra $g(\omega)$ and $\Phi(\omega)$ are related by the equation

$$g(\omega) = \Phi(\omega) F(i\omega) F(-i\omega) = |F(i\omega)|^2 \Phi(\omega) \tag{9.71}$$

where the fact that $F(i\omega)$ and $F(-i\omega)$ are complex conjugates has been used.

Equation (9.71) gives the power spectrum of the output from the power spectrum of the input and the frequency response of the linear system. Even when the frequency response is given in a graph or in a numerical table, the power spectrum $g(\omega)$ can be easily calculated. The usefulness of the concept of transfer function and frequency response is thus again demonstrated. It is of interest to note that since $F(i\omega)$ generally vanishes

for $\omega \to \infty$, the output power spectrum $g(\omega)$ goes to zero for $\omega \to \infty$ faster than the input power spectrum $\Phi(\omega)$. This has the effect of "smoothing" the output function.

9.9 Second-order System. As a simple example, consider the linear system to be of second order. Then the equation of motion is

$$m\frac{d^2y}{dt^2} + c\frac{dy}{dt} + ky = x(t) \tag{9.72}$$

The transfer function $F(s)$ of the system is thus

$$F(s) = \frac{1}{ms^2 + cs + k} = \frac{1}{k}\frac{1}{(s^2/\omega_0^2) + 2\zeta(s/\omega_0) + 1}$$

where ω_0 is the natural frequency of the undamped system and ζ is the ratio of actual damping to the critical damping, defined by Eq. (3.38). Therefore

$$F(i\omega)F(-i\omega) = \frac{1}{k^2\{[(\omega/\omega_0)^2 - 1]^2 + 4\zeta^2(\omega/\omega_0)^2\}}$$

The power spectrum of the output is thus

$$g(\omega) = \frac{\Phi(\omega)}{k^2\{[(\omega/\omega_0)^2 - 1]^2 + 4\zeta^2(\omega/\omega_0)^2\}} \tag{9.73}$$

If we are interested in the mean-square output of the linear system, Eq. (9.21) gives

$$\overline{y^2} = \frac{1}{k^2}\int_0^\infty \frac{\Phi(\omega)\,d\omega}{[(\omega/\omega_0)^2 - 1]^2 + 4\zeta^2(\omega/\omega_0)^2} \tag{9.74}$$

Now if ζ is very small, the denominator of the integrand in Eq. (9.74) is very nearly zero at $\omega = \omega_0$. Therefore, if $\Phi(\omega)$ is a slowly varying function, then

$$\overline{y^2} \approx \frac{\omega_0\Phi(\omega_0)}{k^2}\int_0^\infty \frac{dx}{(x^2 - 1)^2 + 4\zeta^2x^2} = \frac{1}{k^2}\,\omega_0\Phi(\omega_0)\,\frac{\pi}{4\zeta} = \frac{\pi}{2mc}\frac{\Phi(\omega_0)}{\omega_0^2} \tag{9.75}$$

This equation shows that if the damping coefficient c vanishes, the mean-square output will become infinitely large. When c is zero, the transfer function $F(s)$ has the purely imaginary pole $i\omega_0$. This phenomenon of infinite output occurs in general whenever the transfer function of a linear system has a purely imaginary pole. Therefore, for satisfactory operation under random input, the *condition on the system transfer function is that all poles of $F(s)$ should have negative real parts.* This fundamental requirement on the system property is then identical to that for conventional input functions.

Other improvements on the output behavior can generally be accom-

plished by further modification of the system transfer function. For instance, it is quite conceivable that the function $\Phi(\omega_0)/\omega_0^2$ in Eq. (9.75) has a minimum at a reasonable frequency ω_0^* as shown in Fig. 9.6. Then it will be possible to reduce the output random fluctuation by making the system operate effectively at ω_0^*. This

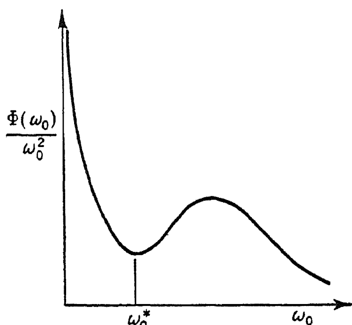


FIG. 9.6

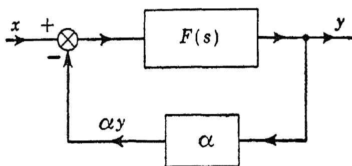


FIG. 9.7

can be accomplished by a simple proportional servo feedback, indicated in Fig. 9.7. Then instead of Eq. (9.72) we have

$$m\,\frac{d^2y}{dt^2} + c\,\frac{dy}{dt} + ky = x - \alpha y$$

or

$$m\,\frac{d^2y}{dt^2} + c\,\frac{dy}{dt} + (k + \alpha)y = x \tag{9.76}$$

The natural frequency of the system can then be made to be ω_0^* by requiring

$$\omega_0^{*2} = \frac{k + \alpha}{m} \tag{9.77}$$

Therefore, by proper choice of α, the output can be reduced.

9.10 Lift on a Two-dimensional Airfoil in an Incompressible Turbulent Flow. As a second example, consider a flat, thin airfoil of chord c moving with a constant velocity U through turbulent air. Let x lie along the chord, z along the span of the wing, and y normal to the span and cord. The components of the fluctuating turbulent velocity u, v, and w are assumed to be small in comparison to U. Because of these turbulent fluctuations, a time-dependent apparent angle of attack α exists at the airfoil, and, hence, fluctuating lift forces are produced. The fluctuating angle of attack α is given by

$$\alpha = \frac{v}{U}$$

as long as the fluctuations are small. $\alpha(t)$ now plays the role of the forcing function. The "response" is the fluctuating lift of the airfoil, or, better, the lift coefficient $C_l(t)$. This is a problem studied by H. W. Liepmann.[1]

[1] H. W. Liepmann, *J. Aeronaut. Sci.*, **19**, 793–801 (1952).

To find the mean square $\overline{C_l^2(t)}$ of the lift coefficient, it is first necessary to define a transfer function for the airfoil. This has already been done in Sec. 3.7. In fact the frequency response $F(i\omega)$, with v as the input and the lift coefficient C_l as the output, is specified by Eqs. (3.54) to (3.58).

Turbulent fluctuations are, however, essentially three-dimensional—that is, u, v, and w will be functions of x, y, z, and t. For the first analysis, it seems sufficient to consider only the component v and its dependence upon x and t. Thus, in turbulent flow we consider a fluctuating velocity or angle of attack of the form

$$\alpha(x,t) = \frac{v(x,t)}{U}$$

If it is now assumed that the turbulence pattern does not change appreciably during a time of the order c/U, then the turbulent angle of attack will also depend upon $t - (x/U)$ only, and Sear's result, given in Sec. 3.7, can be applied. This assumption is frequently made in turbulence analysis and requires essentially the condition that the rate of change of fluid velocity by following a fluid particle is small compared with the rate of change of the fluid velocity at a fixed position. With this assumption,

$$\overline{C_l^2} = 4\pi^2 \int_0^\infty \Phi(\omega)|\varphi(k)|^2 \, d\omega \tag{9.78}$$

where $\Phi(\omega)$ is the power spectrum of v/U.

According to Eqs. (9.31) and (9.32),

$$\Phi(\omega) = \frac{\overline{v^2}}{U^2}\frac{L}{\pi U}\frac{1 + 3\,(L^2\omega^2/U^2)}{[1 + (L^2\omega^2/U^2)]^2} \tag{9.79}$$

Furthermore, Liepmann has discovered that $|\varphi(k)|^2$ can be approximated by

$$|\varphi(k)|^2 \approx \frac{1}{1 + 2\pi k} \tag{9.80}$$

Thus

$$\begin{aligned}\overline{C_l^2} &= 4\pi^2 \frac{\overline{v^2}}{U^2} \int_0^\infty \frac{1 + 3u^2}{(1 + u^2)^2}\frac{1}{1 + \eta u}\, du \\ &= 4\pi^2 \frac{\overline{v^2}}{U^2}\left[\frac{4\eta - \pi}{2\pi(\eta^2 + 1)} + \frac{\eta^2 + 3}{2\pi(\eta^2 + 1)^2}(\eta \log \eta^2 + \pi)\right]\end{aligned} \tag{9.81}$$

where

$$\eta = \frac{\pi c}{L} \tag{9.82}$$

This dependence of the mean-square lift coefficient upon the parameter η is shown in Fig. 9.8.

Evidently, if $c/L \to 0$, we have an airfoil of small chord in a large-scale turbulence, and

$$\overline{C_l^2} \to 4\pi^2 \frac{\overline{v^2}}{U^2} = 4\pi^2\overline{\alpha^2}$$

The airfoil behaves in a quasi-stationary manner with a lift slope of 2π. If, on the other hand, c/L becomes large, we deal with an airfoil of long chord in a small-scale turbulence. It follows from Eq. (9.80) that $\overline{C_l^2} \to 0$. That is to say the "gusts" cancel each other out completely, and the net lift is zero, as might be expected.

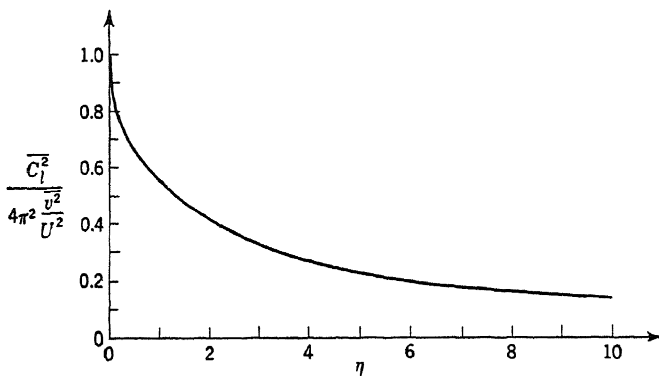


FIG. 9.8

9.11 Intermittent Input. A phenomenon of major importance for aerodynamic buffeting is the so-called "intermittency" in wake flow. That is, the edge of any wake fluctuates with a large-scale motion so that a point situated near the edge will sometimes be within the wake and sometimes outside. For a tail surface situated near the edge of the wake of a stalled or partially stalled wing, this "intermittency" may thus be extremely important in determining the lift forces on the tail wing. For a crude idea of the effect, consider the flow at the tail as a region of uniform down-wash switched on and off at irregular intervals. Such a flow is probably a good model for the conditions in the wake of an intermittently stalling wing. If the probability of switching over from one region to the other is assumed to be governed by Poisson's distribution, one can then apply the power spectrum of Eq. (9.50) with some modifications. The mean deviation is not unity but the mean angle $\sqrt{\overline{v^2}}/U$; and the average time during which the forcing function is switched on is the mean interval $\bar{T}$. Thus the power spectrum is

$$\Phi(\omega) = \frac{\overline{v^2}}{U^2}\frac{\bar{T}}{\pi}\frac{1}{1 + (\omega\bar{T}/2)^2} \tag{9.83}$$

Then the mean-square lift coefficient is approximately

$$\overline{C_l^2} = \frac{\overline{v^2}}{U^2}\frac{\bar{T}}{\pi}4\pi^2 \int_0^\infty \frac{d\omega}{[1 + (\omega\bar{T}/2)^2][1 + \pi(\omega c/U)]}$$

$$= 4\pi^2 \frac{\overline{v^2}}{U^2}\frac{2}{\pi}\frac{\eta \log \eta + \frac{\pi}{2}}{1 + \eta^2} \qquad \text{for } \eta = \frac{2\pi c}{U\bar{T}} \tag{9.84}$$

This relation is plotted in Fig. 9.9. The limiting values at $\eta = 0$ and $\eta \to \infty$ are, of course, the same for the case studied in the previous section.

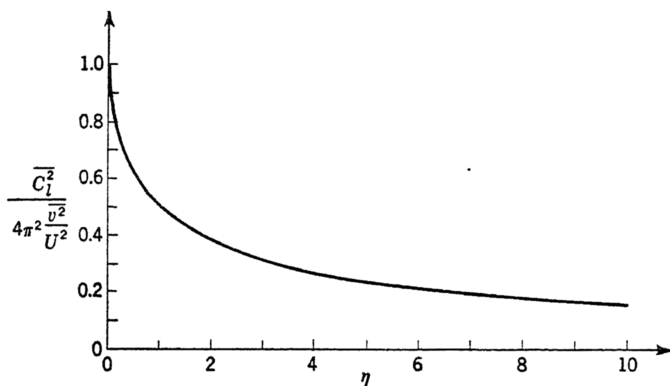


FIG. 9.9

9.12 Servo Design for Random Input. We have already indicated, in connection with the response of a second-order system to random input, the possibility of improving the behavior of the system by servo control. In that instance, however, the feedback mechanism is rather primitive in that the feedback control force required is of the same order of magnitude as the input forcing function. In a more practical design, this feedback mechanism can be made much more subtle, so as to reduce the control force required. For instance, a wing in a turbulent flow can be controlled by a feedback servo which moves the hinged flap. The force necessary to move the flap could be quite small in comparison with the aerodynamic effects such movement produces. The servo link can be thought of as that indicated in Fig. 9.10. The input random function is the turbulent air stream. The first aerodynamic transfer function $F_1(s)$ is the relation between the turbulent air stream and the aerodynamic lift due to the turbulent air stream. This function is approximated by that of Eq. (3.56). As a result of the changing lift and moment, the wing is subject to vertical and torsional motion. The relation between the aerodynamic forces and the wing motion is given by the structural transfer function $F_2(s)$. The wing motion will have two effects. There

are the aerodynamic forces due to wing motion through the second aerodynamic transfer function $F_3(s)$. This is the first feedback loop. This loop, however, is not under designer's control. The designer's control is on the second loop. The wing motion can be used to generate flap motion through the transfer function $F_4(s)$. The flap motion will again

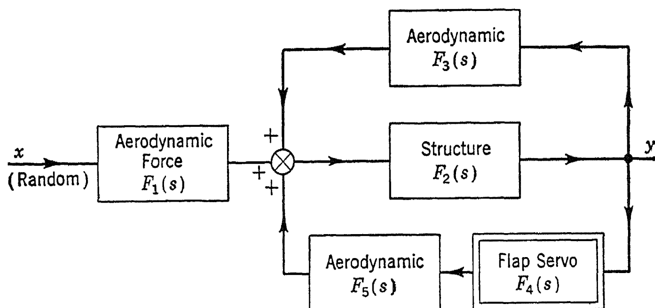


FIG. 9.10

generate aerodynamic forces through the transfer function $F_5(s)$. Thus the input and output relation is as follows:

$$Y = F_2(s)[F_1(s)X + F_3(s)Y + F_5(s)F_4(s)Y]$$

or

$$\frac{Y}{X} = F_s(s) = \frac{F_1(s)F_2(s)}{1 - F_2(s)[F_3(s) + F_5(s)F_4(s)]} \tag{9.85}$$

The over-all system transfer function $F_s(s)$ can thus be modified by changing the servo transfer function $F_4(s)$.

If $\Phi(\omega)$ is the power spectrum of the input x, the power spectrum $g(\omega)$ of the output, or the wing displacement $y(t)$, is then, according to Eq. (9.71)

$$g(\omega) = \Phi(\omega)F_s(i\omega)F_s(-i\omega) \tag{9.86}$$

where $F_s(s)$ is given by Eq. (9.85). Now it is quite conceivable that in order to maximize passenger comfort in the airplane, we would wish to make the acceleration $y''(t)$ of the wing as small as possible. This means we have to minimize $\overline{y''^2}$. Equation (9.16) shows that the mean square of y'' can be calculated from the correlation function. But the correlation function can be calculated from the power spectrum by Eq. (9.22). By combining these equations with Eqs. (9.85) and (9.86), we have

$$\overline{y''^2(t)} = \int_0^\infty \omega^4\Phi(\omega) \left| \frac{F_1(i\omega)F_2(i\omega)}{1 - F_2(i\omega)[F_3(i\omega) + F_5(i\omega)F_4(i\omega)]} \right|^2 d\omega \tag{9.87}$$

The minimization is to be carried out by modifying the servo transfer function $F_4(s)$. To do so we can construct the servo transfer function

with unspecified parameters. Then with all other transfer functions $F_1(s)$, $F_2(s)$, $F_3(s)$, and $F_5(s)$ fixed and the input power spectrum given, say, as Eq. (9.79), $\overline{y''^2}(t)$ can be computed as a function of these unspecified parameters. The problem of minimization is then an ordinary minimum problem with respect to these parameters, and the condition of minimization determines the parameters. The resultant servo transfer function is then the best transfer function for the purpose of maximizing passenger comfort.

The above discussion is but one example of optimum servo design for a specified purpose and specified input. As another instance, the design condition could be the minimization of the mean square of the elastic stress induced in the wing structure by the turbulent air stream. Then the system transfer function will be a different one, but the general formulation of the problem remains the same. Such a problem of quantitative optimum design can be considered as one step beyond the mere requirement of stability and other qualitative criteria of servomechanisms introduced in the previous chapters. This general concept was perhaps first formulated by A. S. Boksenbom and D. Novik.[1] We shall touch upon this problem again in Chap. 16.

[1] A. S. Boksenbom, D. Novik, *NACA TN* 2939 (1953).

CHAPTER 10

RELAY SERVOMECHANISMS

If there is an on-off relay in the servomechanism, the system is called a relay servomechanism. As pointed out in Sec. 6.3, the one great advantage of a relay servomechanism is its low cost. However, since the output of a relay is not proportional to the input, *i.e.*, the input-output relation is not linear, the behavior of a relay servomechanism cannot be analyzed by a linear theory. In this chapter, we shall first present an approximate theory for investigating the stability of relay and other similar servomechanisms, based again on a modification of the Nyquist criterion. In the latter part of this chapter, the more advanced and more difficult problem of how to utilize the inherent nonlinear characteristics of a relay to achieve superior performance from the servomechanism will be discussed. Unfortunately, this particular subject is still far from being completely investigated; the general solution is yet to come.

10.1 Approximate Frequency Response of a Relay. Let us consider a sinusoidal input $x(t)$ of frequency ω and amplitude a,

$$x(t) = a \sin \omega t \tag{10.1}$$

The characteristics of the relay will be idealized in that no time delay is considered, and the action is considered to be instantaneous, *i.e.*, there is no time lag. On the other hand, the hysteresis of the relay action is included: When the input is positive and increasing, the relay output changes from zero to the full value of A at $x = b$. When the input is positive but decreasing, the relay output changes from A to zero at $x = c$. b is greater than c. b is called the pull-in current and c the drop-out current. Pull-in and drop-out for negative input occur at $x = -b$ and $x = -c$, respectively. The input-output relation can then be presented as that of Fig. 10.1. It is evident that there is a phase shift between the input and the output. This phase lag θ is

$$\theta = \frac{1}{2}\left[\sin^{-1}\frac{b}{a} - \sin^{-1}\frac{c}{a}\right] \tag{10.2}$$

The output square waves have a time duration of $2\alpha/\omega$, where α is given by

$$\alpha = \frac{\pi}{2} + \theta - \sin^{-1}\frac{b}{a} = \frac{1}{2}\left[\pi - \sin^{-1}\frac{b}{a} - \sin^{-1}\frac{c}{a}\right] \tag{10.3}$$

The period of the output is the same as that of the input and is equal to $2\pi/\omega$.

Now the output $y(t)$ can be analyzed into a Fourier series,

$$y(t) = \sum_{1}^{\infty} a_n \sin [n(\omega t - \theta)] \tag{10.4}$$

The coefficient a_1 of the first harmonic is simply

$$a_1 = \frac{4A}{\pi} \sin \alpha$$

where α is given by Eq. (10.3). In a relay servomechanism, Fig. 10.2, the relay output is the correction signal which actuates the servo. The

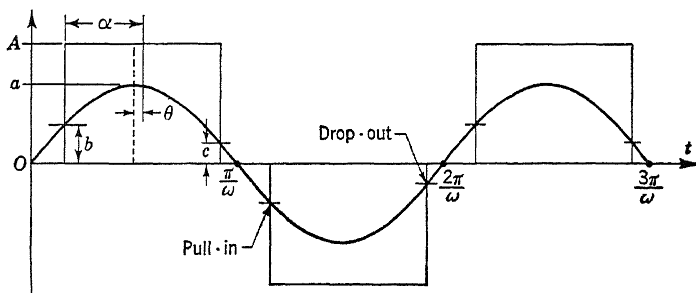


FIG. 10.1

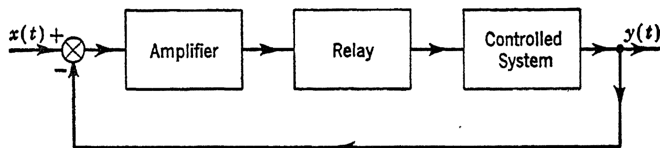


FIG. 10.2

servo has generally the property of a filter and greatly reduces the importance of higher harmonics. Therefore, as an approximation, we may neglect all the higher harmonics and consider the output to be $a_1 \sin (\omega t - \theta)$. Then the ratio of output to input is

$$F_r(i\omega) = \frac{4A \sin \alpha}{\pi a} e^{-i\theta} \tag{10.5}$$

$F_r(i\omega)$ can then be considered the frequency response of the relay. Of course it is not the true frequency response. True frequency response, as defined in the previous chapters, is a function of frequency only, not a function of the input amplitude. The frequency response of Eq. (10.5) is, however, a function of the input amplitude a. On the other hand, $F_r(i\omega)$ is not a function of the frequency ω. Therefore the func-

tional notation adopted is not appropriate; but for easy identification we shall keep the notation in spite of this.

When the amplitude a of the input is very large, we have, from Eqs. (10.2), (10.3), and (10.5),

$$F_r(i\omega) = \frac{4A}{\pi}\frac{1}{a} \qquad \text{for } a \to \infty \tag{10.6}$$

When the amplitude a is very very small, the relay will not give any response. The cutoff point occurs at $a = b$. Then

$$\theta = \alpha = \frac{1}{2}\left(\frac{\pi}{2} - \sin^{-1}\frac{c}{b}\right) \qquad \text{for } a = b \tag{10.7}$$

These limiting values of the relay frequency response are thus completely determined by the relay characteristics.

10.2 Method of Kochenburger. Let us assume for the moment that the amplitudes a of the harmonic components of the input to the relay are all equal. Then the frequency response $F_r(i\omega)$ of the relay is a complex constant, given by Eq. (10.5). If $F_1(i\omega)$ is the frequency response of the rest of the circuit in Fig. 10.2, then the over-all frequency response of the forward circuit is $F_r(i\omega)F_1(i\omega)$. If we apply the Nyquist criterion of Sec. 4.3, then for stability the curve traced on the complex plane by $1/[F_r(i\omega)F_1(i\omega)]$ as ω varies from 0 to ∞ must "enclose" the point -1. In other words the curve $1/[F_r(i\omega)F_1(i\omega)]$ must cross the real axis to the left of the point -1. But for constant input amplitude a to the relay, $F_r(i\omega)$ is a constant, and the above-stated condition for stability is equivalent to requiring the frequency-response curve $1/F_1(i\omega)$ to enclose the point $-F_r(i\omega)$, as ω varies from 0 to ∞. This is the basis of the frequency-response method of Kochenburger[1] for determining the stability of a relay servomechanism. Dutilh[2] developed a similar method independently.

Kochenburger argues that when the amplitudes a of the harmonic components of the input to the relay are not equal, all that is necessary is to apply the stability condition deduced in the previous paragraph to all values of $F_r(i\omega)$ as the amplitude a varies from 0 to ∞. The locus of $-F_r(i\omega)$ for various a is a curve, starting at the cutoff point specified by Eq. (10.7) and ending at the origin of the complex plane. This is shown in Fig. 10.3. Kochenburger's sufficient condition for stability is then: The curve $1/F_1(i\omega)$ must enclose the complete locus of $-F_r(i\omega)$, as shown in Fig. 10.3. Figure 10.4 shows the case of complete instability. The arrows on the $-F_r(i\omega)$ curve show the direction of increasing amplitude of the input to the relay. The arrows on the $1/F_1(i\omega)$ curve show the direction of increasing frequency, starting at the origin with $\omega = 0$.

[1] R. J. Kochenburger, *Trans. AIEE*, **69**, 270–284 (1950).

[2] J. R. Dutilh, *L'Onde électrique*, **30**, 438–445 (1950).

Besides these cases of complete stability and complete instability, there are various cases of partial stability or instability, with the possibility of persistent oscillations of a certain frequency at constant amplitude. For instance, Fig. 10.5 shows a case having a convergent point. For small amplitudes, the $-F_r(i\omega)$ points are "outside" the $1/F_1(i\omega)$ curve, and thus the system is unstable. Then the amplitude of the oscillations will

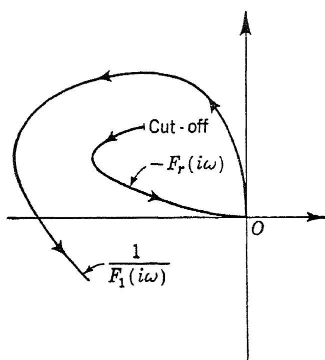


Fig. 10.3

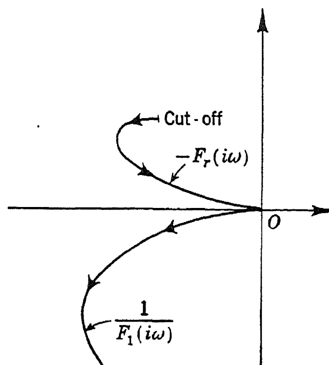


Fig. 10.4

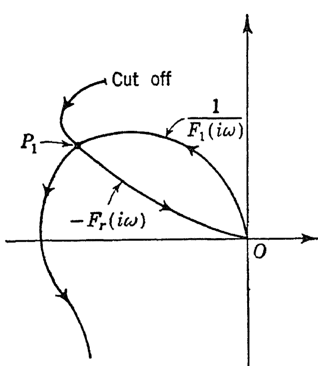


Fig. 10.5

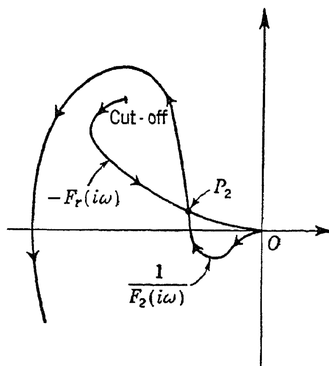


Fig. 10.6

increase. As the amplitude increases, the point $-F_r(i\omega)$ moves towards the curve $1/F_1(i\omega)$, and finally the point P_1 is reached. Then the system will oscillate with a frequency and an amplitude corresponding to that point. The oscillation is thus self-starting, and the behavior is called *"soft" self-excitation.* There is no tendency to move away from P_1, because any increase in amplitude will meet damping effects by entering the stable region of the diagram. P_1 is thus a convergent point, and the system will always oscillate. Figure 10.6 shows a different case where the point P_2 of the intersection of the $-F_r(i\omega)$ curve and $1/F_1(i\omega)$ curve

is a divergent point. The system always tends to move away from that point. It is, however, stable for small disturbances.

Figures 10.7 and 10.8 show yet more complicated cases, having both convergent points P_1 and divergent points P_2. The system of Fig. 10.7 will oscillate with disturbances of sufficiently large amplitude. This behavior is called *"hard" self-excitation.* The system of Fig. 10.8 will always oscillate unless the amplitude of the disturbance is too large. For very large disturbances, the system will diverge. All the cases depicted in Figs. 10.5 to 10.8 show the peculiar dependence of the behavior of the system on the amplitude of the disturbances, and the possibility of persistent oscillation at fixed frequency *and* amplitude. These are all

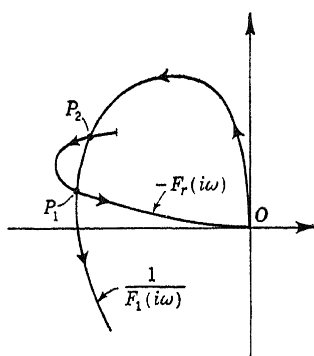


Fig. 10.7

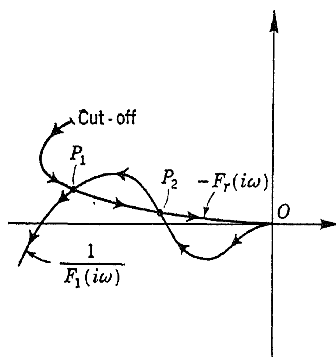


Fig. 10.8

characteristics of nonlinear systems and are not present in the linear systems studied in the previous chapters. Such behaviors are of course to be expected from our introductory discussions of nonlinear systems in Sec. 1.3.

10.3 Other Frequency-insensitive Nonlinear Devices. The Kochenburger method is a very effective solution for the problem of stability of a relay servomechanism. It can be applied to systems with a considerable degree of complexity; and it allows the direct inclusion of experimentally measured frequency-response data. For most applications where the servo after the relay gives sufficient filtering action, the neglection of higher harmonics in the relay output is fully justified. The prediction of theory is found to be in full agreement with experimental observation. Therefore, if the only design criterion is stability, then the Kochenburger method solves the problem completely for relay servomechanisms.

In fact, the Kochenburger method is not only applicable to relay servomechanisms, but can be applied equally well to many other nonlinear devices. The essential point of this method of analysis is the discovery

that the behavior of the relay is frequency invariant but amplitude variant, while the behavior of a linear system is frequency variant but amplitude invariant. Now there are many nonlinear devices which have the same behavior as the relay. For instance, a gear train with backlash is such a device. We can see this in the following way: Let θ_1 be the angular position of the shaft of the driving motor, which is rigidly connected to the first gear of the train, and let θ_2 be the angular position of the shaft of the last gear of the train. Then the relation between θ_1 and θ_2 can be represented as in Fig. 10.9, where 2δ is the total backlash of the train. If θ_1, the input to the gear train, is a sinusoidal oscillation, θ_2, the output, is a sort of clipped sinusoidal wave with a lag in phase (Fig. 10.10). It is easy to see that, since the relation between θ_1 and θ_2 is not explicitly dependent upon time, the wave form of θ_2 will not be modified by a change in frequency of θ_1. Thus the response of the gear train is amplitude variant but frequency invariant. If we denote the ratio of the amplitude of the fundamental of θ_2 to that of the input θ_1 and the phase lag by the response $F_g(i\omega)$, then $F_g(i\omega)$ is a function of a, but not of ω. This response function can then be used to study servomechanisms containing such gear trains with backlash in exactly the same manner as the relay response function $F_r(i\omega)$ in the preceding section.

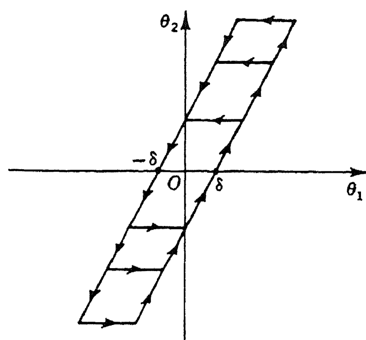


Fig. 10.9

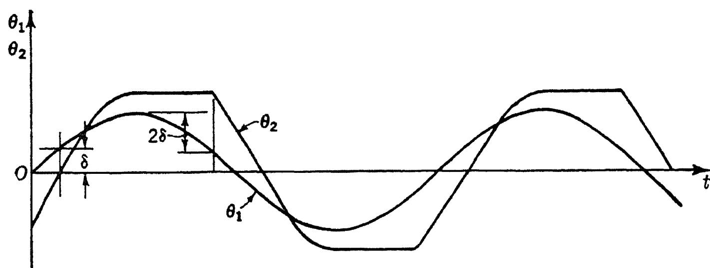


Fig. 10.10

10.4 Optimum Performance of a Relay Servomechanism. It is to be borne in mind that, according to the *Kochenburger diagram* for stability, the curve $1/F_1(i\omega)$ has to avoid the whole locus of the relay response $-F_r(i\omega)$, not just the single point -1 as in the case of a conventional servomechanism. This is a more stringent condition on other parts of the system, and is also the reason why the performance of a relay servo-

mechanism tends to be inferior to the performance of a conventional servomechanism. However, this is not an intrinsic shortcoming of the relay servomechanism. In fact, instead of merely asking for stability, we can consider the relay as a switching device, capable of producing a positive constant correction signal, a negative constant correction signal, or no correction signal, and then ask the question: How should we switch the relay in accordance with the output so as to obtain the optimum performance of the over-all system? For instance, the requirement on performance may be the quickest possible return to the normal state after a disturbance. This requirement not only guarantees the return to normal state (stability) but also specifies the speediest return. The solution to this problem of optimum performance is to specify the proper switching action of the relay as a function of the output, and this *switching function* is the basis for the actual design of the servo system. A relay servomechanism designed with this more general point of view will certainly have a performance superior to that of a conventional linear servomechanism, because the nonlinear characteristics of the relay are utilized to the fullest extent.

10.5 Phase Plane. If y is the output and x the input, the differential equation of a general second-order system, linear or nonlinear, can be written as

$$f(y,\dot{y},\ddot{y};t) = x(t) \tag{10.8}$$

where the dot denotes differentiation with respect to time. We can rewrite Eq. (10.8) as the system

$$\left.\begin{aligned} f\left(y,\dot{y}\,\frac{d\dot{y}}{dt};t\right) &= x(t) \\ \frac{dy}{dt} &= \dot{y} \end{aligned}\right\} \tag{10.9}$$

If we consider y and $\dot{y}$ as the dependent variables, Eq. (10.9) is a system of two simultaneous first-order differential equations for the unknowns y and $\dot{y}$. If the input is absent, $x = 0$, and if f is not a function of t, that is, the system is *autonomous*, as is frequently the case, then the first equation of Eq. (10.9) can be solved for $d\dot{y}/dt$ and gives $d\dot{y}/dt$ as a function of y and $\dot{y}$. Then the system can be written as

$$\left.\begin{aligned} \frac{d\dot{y}}{dt} &= \dot{y}k(y,\dot{y}) \\ \frac{dy}{dt} &= \dot{y} \end{aligned}\right\} \tag{10.10}$$

This system of equations does not contain t explicitly. By dividing the first equation of Eq. (10.10) by the second, we have

$$\frac{d\dot{y}}{dy} = k(y,\dot{y}) \tag{10.11}$$

This is now a first-order equation, with y as the independent variable and $\dot{y}$ as the dependent variable. After this equation is solved, Eq. (10.10) can be used to calculate the relation between t and y.

Physically, the procedure outlined in the previous paragraph is based upon the concept of characterizing the state of the system by y and $\dot{y}$, instead of the more conventional method of specifying it by y and t. If y is the variable describing the position of a point mass, then $\dot{y}$ is the "velocity." $\dot{y}$ can thus be considered as representative of the momentum of the point mass. y and $\dot{y}$ then represent the position and the momentum, respectively, of the point mass. Physicists call such a representation of state the representation in *phase space*. In the particular case discussed, the phase space has only two dimensions; it is thus a *phase plane*. The behavior of the second-order system is then specified by a curve in the phase plane. Every point on the curve represents the state of the system at a certain time t. It is customary to indicate the sense of increasing time along the curve by an arrow, as Fig. 10.11. If the order n of the system is higher than 2, the phase space will be of n dimensions, and the behavior of the system is represented by a curve in this n-dimensional space.

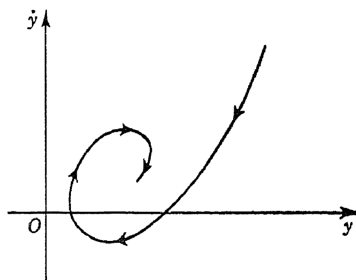


Fig. 10.11

The practical advantage of phase-plane representation is that a very large number of nonlinear systems of second order are autonomous systems and can be expressed as Eq. (10.11), and this equation can be solved, at least graphically, by the isocline method. In fact, the character of the system is at once clear when the *field* of directions of the curves, as specified by Eq. (10.11), is indicated in the phase plane. The use of such geometrical properties of the phase plane is at the heart of the theory of nonlinear oscillations and is called the topological methods of nonlinear mechanics.

To translate our previous concepts to the language of the phase plane, let us consider the simple problem of a linear second-order system without a forcing function,

$$\frac{d^2y}{dt^2} + 2\zeta\frac{dy}{dt} + y = 0 \tag{10.12}$$

which is obtained from Eq. (3.39) by choosing the time unit in such a way as to make the natural frequency ω_0 unity. ζ is of course the ratio

of the damping of the system to critical damping. For oscillations, $|\zeta| < 1$. Equation (10.12) can be rewritten as

$$\left.\begin{aligned} \frac{d\dot{y}}{dt} &= -2\zeta\dot{y} - y \\ \frac{dy}{dt} &= \dot{y} \end{aligned}\right\} \tag{10.13}$$

Therefore, corresponding to Eq. (10.11), we have

$$\frac{d\dot{y}}{dy} = -\frac{2\zeta\dot{y} + y}{\dot{y}} = -2\zeta - \frac{y}{\dot{y}} \tag{10.14}$$

It is clear then that the lines of constant slope $d\dot{y}/dy$ are radial lines from the origin of the phase plane. Figures 10.12 to 10.16 are the five cases

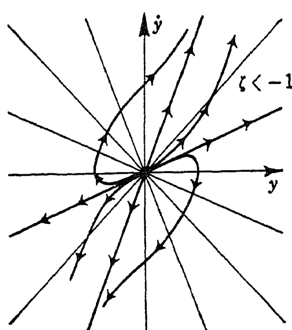


Fig. 10.12

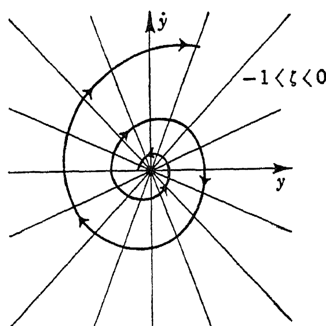


Fig. 10.13

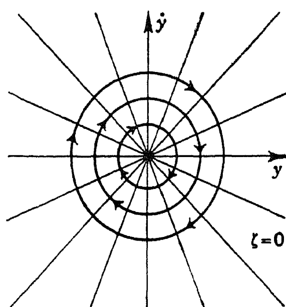


Fig. 10.14

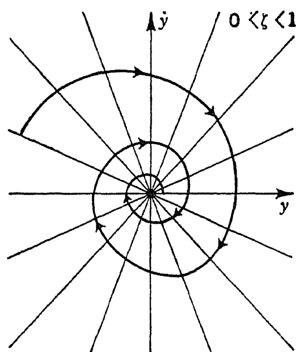


Fig. 10.15

with $\zeta < -1$, $-1 < \zeta < 0$, $\zeta = 0$, $0 < \zeta < 1$, and, finally, $1 < \zeta$. Figures 10.12 and 10.16 are the nonoscillatory cases, Figs. 10.13 and 10.15 are the oscillatory cases, and Fig. 10.14 is the case of purely harmonic oscillation.

In the above figures, the origin of the phase plane corresponds to the

equilibrium state, where both dy/dt and $d\dot{y}/dt$ vanish. Mathematically, the origin is the *singular point* of the system of equations of Eq. (10.13). The character of the equilibrium state is, however, quite different for the three different cases $\zeta < 0$, $\zeta = 0$, and $0 < \zeta$. Figures 10.12 and 10.13 show that when $\zeta < 0$, the lines of behavior of the system all diverge from the equilibrium state. Therefore the origin is an *unstable equilibrium point.* Figures 10.15 and 10.16 show that when $0 < \zeta$, the lines of behavior all converge to the equilibrium state. Then the origin is a *stable equilibrium point.* Mathematically, the origin in Figs. 10.12 and 10.16 is a point through which all lines of behavior pass and is called the *node.* The origin in Figs. 10.13 and 10.15 is the center of the spirals and is called the focus. In the special case of Fig. 10.14, where $\zeta = 0$, the lines of behavior circle the origin. The origin is then called the *center.*

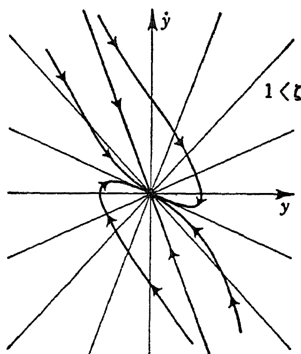


FIG. 10.16

If the basic equation of the second-order system has a constant forcing term, *i.e.*,

$$\frac{d^2y}{dt^2} + 2\zeta\frac{dy}{dt} + y = c \tag{10.15}$$

where c is a constant, then

$$\frac{d^2(y - c)}{dt^2} + 2\zeta\frac{d(y - c)}{dt} + (y - c) = 0$$

Therefore the lines of behavior in the phase plane are entirely similar to those indicated in Figs. 10.12 to 10.16, with only the modification of translating the equilibrium point to $y = c$, on the y axis.

10.6 Linear Switching. In the subsequent discussion, we shall simplify the problem of switching the relay by assuming only two states for the relay: unit positive output and unit negative output. The specification of unit output is evidently not a restriction to our problem. But before attacking the optimum switching problem proper, let us consider the simpler case of *linear switching, i.e.*, the case where the forcing function c generated by the relay is equal to unity but has the same sign as $ay + b\dot{y}$. The purpose here is to demonstrate some of the characteristics of the problem.

The performance of a linearly switched relay servomechanism is analyzed by Flügge-Lotz,[1] and Flügge-Lotz and Klotter.[2] The following

[1] I. Flügge-Lotz, *Z. angew. Math. Mech.*, **25–27**, 97–113 (1947).
[2] I. Flügge-Lotz and K. Klotter, *Z. angew. Math. Mech.*, **28**, 317–337 (1948).

discussion is a summary of the qualitative results of their investigation. The differential equation of the system studied by the above authors is

$$\frac{d^2y}{dt^2} + 2\zeta \frac{dy}{dt} + y = \operatorname{sgn}(ay + b\dot{y}) \qquad \text{for } 0 < \zeta < 1 \qquad (10.16)$$

A curve in the phase plane with unit positive forcing function is called a *P arc*. A curve with unit negative forcing function is called an *N arc*. The system of all *P arcs* is called the *P system*, and the system of all *N* arcs is called the *N system*. It is clear from our discussion in connection with Eq. (10.15) that for the particular equation under study, the *P* system is a system of converging spirals with its focus at the point $y = +1, \dot{y} = 0$; and that the *N* system is a system of converging spirals with its focus at the point $y = -1, \dot{y} = 0$. The desired end state of the system is, of course, the origin $y = \dot{y} = 0$.

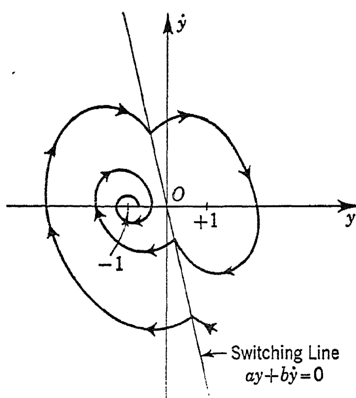


FIG. 10.17

According to the signs of a and b in the switching function of Eq. (10.16), four cases can be defined. Let us call *Case* 1 the case with $a > b$, $b > 0$. Then the switching line $ay + b\dot{y} = 0$ is a straight line passing through the origin of the phase plane and lies in the second and the fourth quadrants. To the right of it, the sign of $ay + b\dot{y}$ is positive, and the region is that of a *P* system. To the left of it, the sign of $ay + b\dot{y}$ is negative, and the region is that of an *N* system. At the switching line, the two systems join, and there the corners of the curves of behavior occur (Fig. 10.17). The condition for the existence of a periodic solution is that there should exist a *P* arc whose intersections with the switching line are equidistant from the origin; for then, by symmetry, there exists an *N* arc on the other side of the switching line joining the same two points, and these two arcs together form a closed curve in the phase plane. Such periodic solutions are called limit cycles in the terminology of nonlinear mechanics. We shall see presently that a periodic solution can occur in our case.

Let S_P and S_N be the points on the switching line where a *P* and an *N* curve, respectively, are tangent; let R_P and R_N be the last intersections preceding S_P and S_N with the switching line of the *P* and *N* curves through these points. S_P and S_N are symmetrical with respect to the origin, as are R_P and R_N. Suppose that ζ, a, and b are such that R_N is outside the segment S_PS_N, as shown in Fig. 10.18. A solution starting

sufficiently near the segment S_PS_N will move away from the line in one direction or the other, according to the side of the line on which its initial point lies, and never return to the switching line at all. It can also be shown that every P arc which lies to the right of the switching line and has its end on this line also has the property that its terminal point is nearer the origin than its initial point; hence the condition for a periodic solution can never be satisfied. The lines of behavior of the system then always spiral to one of the foci, and the end state is not the origin of the phase plane.

If, however, R_N and R_P are on the segment S_PS_N (Fig. 10.19), then there exists a periodic solution. The reason is as follows. The P arc

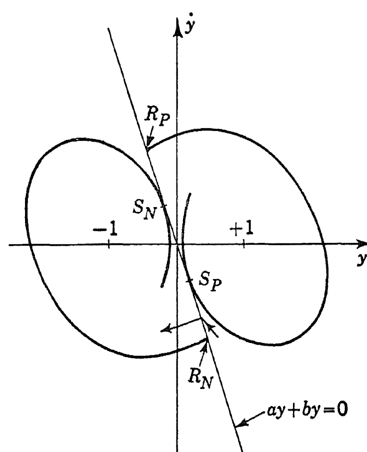


Fig. 10.18

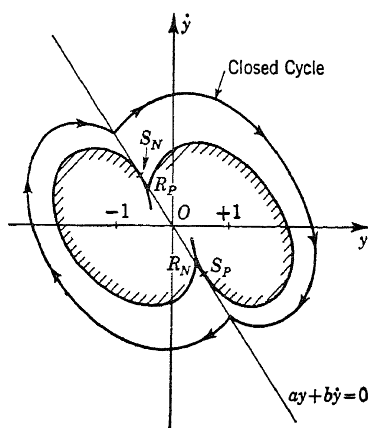


Fig. 10.19

R_PS_P starts at a point which is nearer to the origin than its end point, according to our assumption. However, at large distances from the origin, where the effects of shifting the foci of the converging spirals from the origin to +1 and −1 for the P system and the N system, respectively, are negligible, a P arc must start on the switching line at a point farther away from the origin than its end point on the switching line. These arcs then have the reverse property of R_PS_P. Therefore, by continuity, some intermediate P arc of this type must begin and end at the same distance from the origin. This is the condition for the existence of a periodic solution, which is shown in Fig. 10.19 as a closed cycle. Extended analysis bears this out and shows that the periodic solution is unique and, more important, *orbitally stable:* All solutions beginning outside the periodic solution spiral onto it, and those solutions which begin inside the periodic solution but outside the area enclosed by $R_PS_PR_NS_N$, the shaded area in Fig. 10.19, also spiral onto it. Lines of

behavior beginning within the shaded area will spiral to one of the foci. Here again we have no possibility of reaching the origin.

Case 2 is the case where $a > 0$, but $b < 0$. Now the switching line $ay + b\dot{y} = 0$ passes through the first and third quadrants. The P system and the N system are the same as in the previous case. No periodic solution occurs in this case. Let R_P, R_N, S_P, and S_N be defined as before; then the points R_P, S_P, 0, S_N, and R_N lie on the switching line in this order, as shown in Fig. 10.20. But on the interval S_PS_N a new phenomenon occurs. Consider any solution which reaches this interval, say at the point E. What does the solution do at this point? It should proceed along an N arc, because switching occurs on this line. However, the N arc from E goes back into the same half plane from which the solution

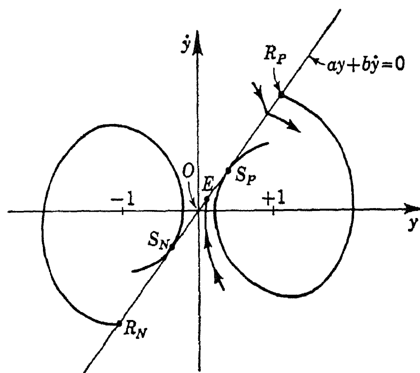


FIG. 10.20

entered E, and on this side a solution can contain only P arcs; on the other hand, the solution certainly cannot follow the P arc through E beyond this point. We may say then that the solution is not defined beyond E, it "ends" at E. Any solution starting outside the region R_PR_N spirals in toward the origin until it reaches the interval R_NR_P. If it reaches the interval R_PS_P, it will spiral to $+1$. If it reaches R_NS_N it will spiral to -1. If it reaches S_PS_N, we have the curious phenomenon of "ending" the solution.

In reality, the behavior of the system cannot "end" but must go on. The paradox is resolved by the observation that switching action always has a *time lag*, and when a solution meets the switching line, it actually proceeds for some distance beyond it before it has the change of sign of the forcing function. In Case 1 such a time lag, provided that it is not too large, does not affect the essential behavior of the system. But in the present case, the time lag avoids the necessary "end" of the solution. Consider a solution entering an end point. Because of the time lag, it no longer ends there, but proceeds for a certain distance beyond;

then switching occurs, and the solution in the phase plane makes a "corner," where the solution is still defined. From this corner, it crosses the switching line in the reverse direction, moves for a short distance beyond, has another corner, and so on, as shown in Fig. 10.21. From that figure, it is also seen that such zigzag action of the system results in its creeping out of the region S_PS_N, and eventually the solution will

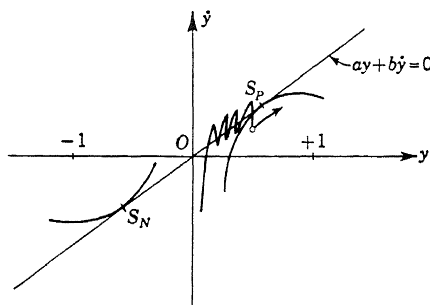


Fig. 10.21

spiral into one of the foci. Time lag will thus avoid the ending of the solution, but still the behavior of the system is unsatisfactory because the solution is not able to reach the origin.

Case 3 is the case where $a < 0$, $b > 0$. In this case the switching line lies again in the first and the third quadrants of the phase plane. But the P system now lies to the left of the switching line, and the N system now lies to the right of the switching line. In this case, a stable periodic solution, or a stable limit cycle (Fig. 10.22), always exists, and it dominates the whole situation, for all other solutions spiral into it. Here again the origin of the phase plane cannot be reached.

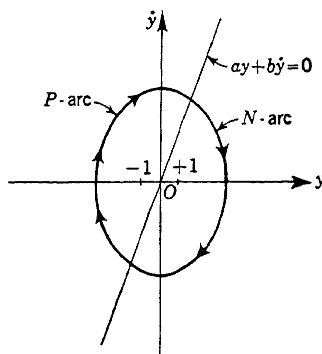


Fig. 10.22

Case 4 is the case where both a and b are negative. The switching line is the same as in Case 1, but the arcs are the same as in Case 3. It may be shown that no periodic solution can exist in this case. Without time lag in the switching action, the segment S_PS_N (Fig. 10.23) consists of end points, and by tracing the solutions which end on it backwards one can see that these cover the entire plane; thus in this case all solutions "end" on the segment S_PS_N of the switching line.

But here again, as in Case 2, the presence of time lag makes a difference. The time lag makes no difference of importance until the solution

in question reaches S_PS_N; then, instead of ending, the solution proceeds for a small distance beyond the switching line, has a corner, crosses the switching line again, has another corner, and so on. It may be seen from Fig. 10.23 that this motion forces the system to the origin. The system will finally oscillate at high frequency and small amplitude around the origin—the smaller the time lag, the higher the frequency. This is what is called "chattering" of the servo system.

It is thus seen that out of the four cases discussed, only *Case* 4 gives a system seeking the desired equilibrium state. But even then, the

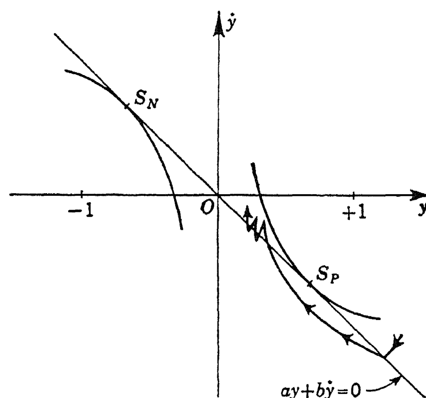


FIG. 10.23

system will chatter near the equilibrium state. Our discussion of the analysis of Flügge-Lotz and Klotter then demonstrates the shortcomings of linear switching. Optimum switching for best performance of the servomechanism is definitely not linear switching. In the following sections, we shall show this.

10.7 Optimum Switching Function. An autonomous second-order system with forcing function of unit magnitude can be written as

$$\left.\begin{aligned} \frac{dy}{dt} &= \dot{y} \\ \frac{d\dot{y}}{dt} + g(y,\dot{y}) &= \varphi(y,\dot{y}) \end{aligned}\right\} \tag{10.17}$$

where $\varphi(y,\dot{y})$ is a discontinuous function with only the two possible values $+1$ or -1. Then the optimum switching problem can be defined as follows: find the function $\varphi(y,\dot{y})$ such that, beginning at any point p of the phase plane, the solution will pass through the origin 0, and the length of time necessary to move from p to 0 along the solution from p is minimal with respect to φ—no other φ could make this time shorter.

Then $\varphi(y,\dot{y})$ is the *optimum switching function.* This particular switching problem was studied by Bushaw,[1] and the special case of linear $g(y,\dot{y})$, that is, $g(y,\dot{y}) = 2\zeta\dot{y} + y$, was completely solved by him for all real values of ζ. The mathematical arguments which led to Bushaw's results are, however, complicated and are difficult to extend to other cases. We shall thus limit ourselves to indicating his solution.

A general result, good for any continuous $g(y,\dot{y})$, is Bushaw's *canonical path.* A path is a line of behavior in the phase plane. Since $\varphi(y,\dot{y})$ can take only the value $+1$ or -1, a path consists of P arcs and N arcs. A junction of these arcs is called a PN corner if with increasing time the switching is from $+1$ to -1 for the forcing function. Similarly, a junction is called an NP corner if the switching is from -1 to $+1$. A path is called canonical if it contains no NP corners above the y axis and no PN corners below the y axis. The importance of the canonical path is that a minimal path, *i.e.*, a path of minimum time, must be canonical.

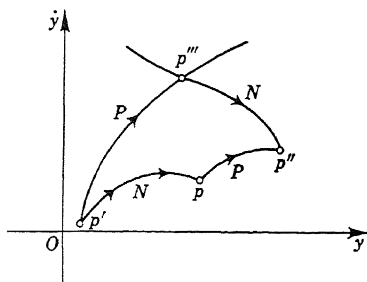


FIG. 10.24

That is, given any path Δ from a point p which is not canonical, one can find a canonical path from p which is shorter than Δ in terms of time. This can be shown quite easily. Given, say, a path with the NP corner p above the y axis, as shown in Fig. 10.24, one denotes by p' either the last corner of the path preceding p or the last intersection preceding p of the path with the y axis, whichever is nearer p, and one denotes by p'' the corresponding point following p. We then draw the P curve forward from p' and the N curve backward from p''. Now, from the basic equations of Eq. (10.17), we have

$$\frac{d\dot{y}}{dy} = \frac{-g(y,\dot{y}) + \varphi(y,\dot{y})}{\dot{y}} \tag{10.18}$$

Therefore, at any point in the phase plane, the slope of the P curve is always greater, algebraically, than the slope of the N curve. Thus the configuration of the paths must be like that indicated in Fig. 10.24. If we now modify the given path by replacing $p'pp''$ by $p'p'''p''$, the NP corner p is removed, and the path is made canonical. If we denote by $t(p'pp'')$ the time interval in going from p', through p to p''; and by $t(p'p'''p'')$, the time interval along the canonical path, then

[1] D. W. Bushaw, Experimental Towing Tank, Stevens Institute of Technology Report 469, Hoboken, N. J. 1953.

$$t(p'pp'') = \int_{p'pp''} \frac{dy}{\dot{y}}$$

$$t(p'p'''p'') = \int_{p'p'''p''} \frac{dy}{\dot{y}}$$

But at any y, the value of $\dot{y}$ on the canonical path is greater than the value of $\dot{y}$ on the original path, and therefore $t(p'p'''p'') < t(p'pp'')$. Thus the canonical path is "shorter" than the noncanonical path.

As a simple example of the application of the theory of the optimum switching function, let us take $g(y,\dot{y}) = \zeta\dot{y}$. Then the system of Eq. (10.17) becomes

$$\left.\begin{aligned} \frac{dy}{dt} &= \dot{y} \\ \frac{d\dot{y}}{dt} &= -\zeta\dot{y} + \varphi(y,\dot{y}) \end{aligned}\right\} \tag{10.19}$$

The P system and the N system of arcs depend, of course, on the value of ζ; but since Eq. (10.19) does not explicitly contain y, these systems of

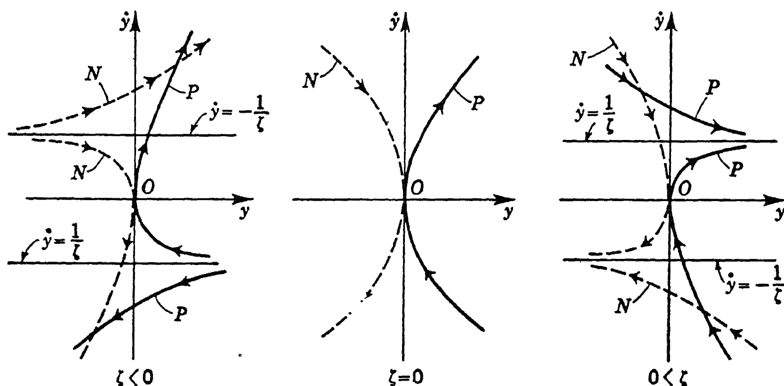


FIG. 10.25

arcs consist of parallel curves shifted along the y axis. A typical P arc and an N arc with one branch passing through the origin, for each of the three possible values of ζ, are shown in Fig. 10.25. The case of $\zeta < 0$ is different from the other cases, in that in order to reach the origin the initial value of $\dot{y}$ must be within the range $-1/\zeta$ to $+1/\zeta$. For this case then, the problem of optimum switching has meaning only if the initial $\dot{y}$ is within this specified range.

We shall denote by Γ that part of the P curve through the origin which lies below the y axis, and by Γ^- its reflection about the origin. Γ^- is therefore that part of the N curve through the origin which lies above the y axis. Γ and Γ^- together gives a curve C. Bushaw showed that C is the *optimum switching line* in the sense that above C the optimum

switching function $\varphi(y,\dot{y})$ should be -1, and below C the optimum switching function $\varphi(y,\dot{y})$ should be $+1$. This is indicated in Fig. 10.26. Physically, the switching is done as follows. From any point p above C, the forcing function should be -1, and the system follows the N arc to the switching line C. There the forcing function changes to $+1$, and the system follows C to the origin. If the initial point p is below C, the forcing function should be $+1$, and the system follows the P arc to the

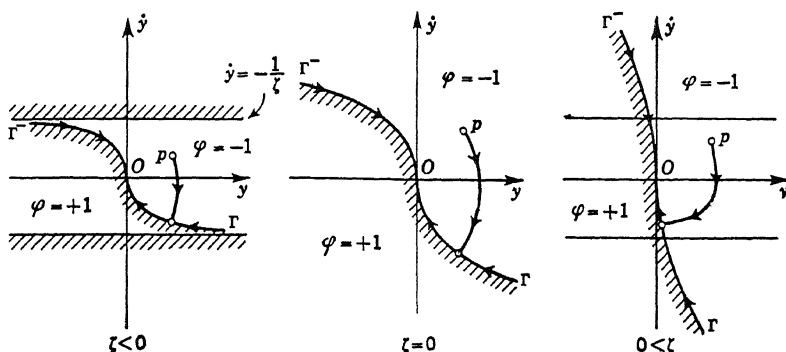


FIG. 10.26

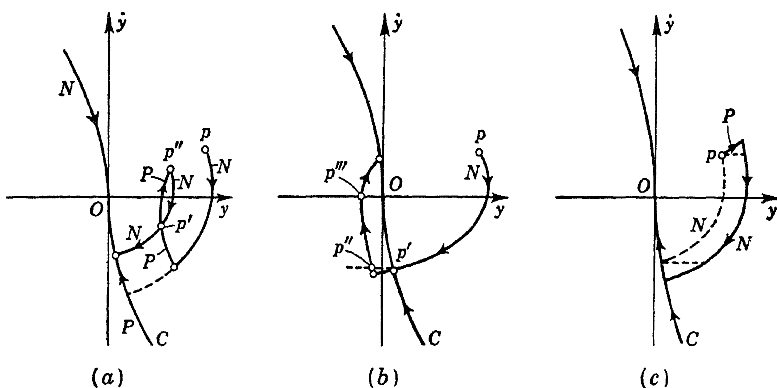


FIG. 10.27

switching line C. There the forcing function changes to -1 and the system follows C to the origin.

That Bushaw's solution for the optimum switching line is correct can be seen as follows. First of all, we note that in order to reach the origin, we must use C for the last part of the path, because C is the only curve through the origin. Now suppose that the initial point is above C and that the optimum path, according to Bushaw, is that indicated in Fig. 10.27(a) by the single N arc from p to C and then the path along C to the origin. Then if the switching is done too early, we have an NP

corner before reaching C. In order to reach C we have to switch again and make a PN corner. If this switching is done at p' when $\dot{y}$ is still negative, then we have a PN corner below the y axis; this violates the rule for a canonical path. The time along the modified path is definitely longer than the optimum path. If the PN corner is made at p'' when $\dot{y}$ is positive, the time required will be even larger, because the path then has a closed loop. Thus switching too early is disadvantageous. Figure 10.27(b) shows the case of switching too late. Since the paths $p'0$ and $p''p'''$ are equivalent and thus require the same time interval, it is easily seen from the figure that switching too late is also detrimental. Figure 10.27(c) shows yet another variation, where the first part of the path is a P arc instead of an N arc. But it is apparent from the figure that this variation is also worse than the optimum path. These considerations indicate the correctness of choosing the canonical path as the optimum switching line.

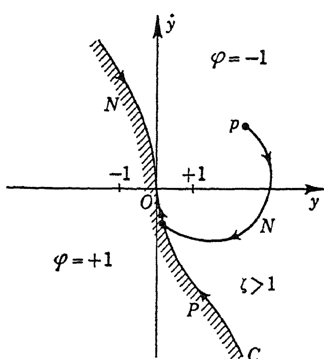


FIG. 10.28

10.8 Optimum Switching Line for Linear Second-order Systems. Bushaw has determined the optimum switching line for the linear second-order systems with $g(y,\dot{y}) = 2\zeta\dot{y} + y$, for all real values of ζ. We shall only state his result here without proof; but in view of our discussion of the simple case in the preceding section, the general character of the result can be easily understood. The P system and the N system for this $g(y,\dot{y})$ are simply the family of curves in Figs. 10.12 to 10.16 with the origin shifted to $(+1,0)$ and $(-1,0)$, respectively.

The case nearest to our simple example is the case of $\zeta > 1$. Then the switching line C consists of the P arc *from* infinity *to* the origin of the phase plane, and the N arc from infinity to the origin. Again, above C, the switching function φ takes the value -1; and below C, φ is equal to $+1$. The optimum path from an initial point above C is thus as indicated in Fig. 10.28.

When $\zeta < -1$, then, as in the simple example, only from points within a limited region of the phase plane can the system reach the origin; because without the forcing function the system is unstable. Bushaw specifies the boundaries of this region as the P arc *from* the point $(+1,0)$ *to* the point $(-1,0)$ and the N arc from the point $(-1,0)$ to the point $(+1,0)$, as shown in Fig. 10.29. The switching problem has meaning only for initial points within this region. The optimum switching line C consists of the P arc *to* the origin and the N arc to the origin. Above C,

the switching function φ is equal to -1, and below C, φ is $+1$. The optimum path from an initial point p above C is shown in Fig. 10.29.

When $\zeta = 0$, the P system and the N system are circles with the centers $(+1,0)$ and $(-1,0)$, respectively. The optimum switching line C (Fig. 10.30) is a series of semicircular arcs of radius unity, starting at the origin, and stretching along the y axis in both directions. The arcs are below the y axis for positive y and are above the y axis for negative y. Above the switching line C, the switching function φ is equal to -1; below it, φ is $+1$. Thus, starting from a point p as shown in Fig. 10.30, the path follows an N arc, or a circular arc with its center on $(-1,0)$. When the path intersects C at a, the path becomes a P arc, or a circular arc with its center on $(+1,0)$, until the path intersects C again at b. At b, the path changes again into an N arc, until the next intersection with the switching line C, etc. The last intersection with C is d, and, from there, the path follows C into the origin. This is a considerably more complex switching performance than that for $\zeta > 1$.

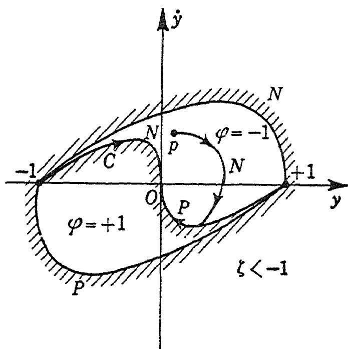


Fig. 10.29

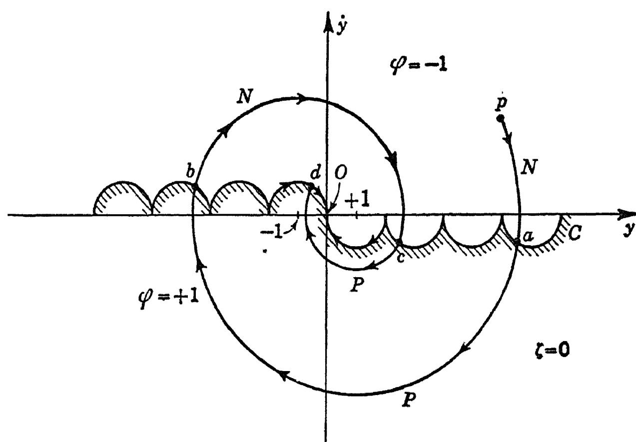


Fig. 10.30

The next more difficult case is the case of converging spirals, or $0 < \zeta < 1$. For this case, Bushaw showed that the optimum switching line should be constructed as follows: We first draw a P spiral backward in time, starting at the origin. The first arc of C is the first arc of P, from the origin to the first intersection with the y axis. Then we draw the

reflection of all arcs above the y axis about their *right* point of intersection with the y axis. Then we assemble all these arcs below the y axis into a continuous curve by moving the consecutive arcs parallel to the y axis until they join end to end, starting from the origin and extending to the right. This is the positive half of the optimum switching line. The negative half of the switching line is then obtained by reflection about the origin. Again, above the switching line C (Fig. 10.31), the switching function φ is -1; below it, φ is $+1$. The situation is then very much the same as for the case $\zeta = 0$ shown in Fig. 10.30; the only difference is the replacement of circular arcs by spiral arcs.

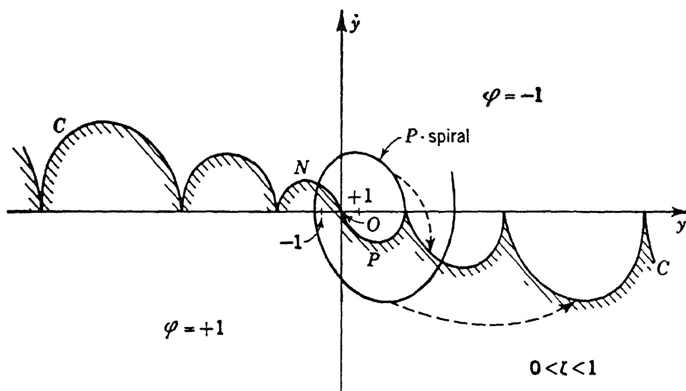


FIG. 10.31

The last case is the case of diverging spirals, or $-1 < \zeta < 0$. The method of construction of the optimum switching line is exactly the same as in the preceding case, except that now the consecutive spiral arcs diminish in size instead of increasing. The length of the switching line is thus finite, spanning the y axis from $(-a,0)$ to $(+a,0)$, as shown in Fig. 10.32. This is as it should be; because here we have negative damping, and, as in the case of Fig. 10.29, the path can reach the origin only if the initial point is within a certain region near the origin. In fact, the boundary of this region consists of the P arc from the point $(a,0)$ to $(-a,0)$ and the N arc from the point $(-a,0)$ to $(a,0)$. The optimum switching function φ takes the value -1 for points above C in the phase plane, and the value $+1$ for points below the switching line.

The boundary curves of the regions of possible optimum switching as indicated in Figs. 10.26 and 10.29 are evidently the limit cycles with switching points at $\dot{y} = 0$. They each then represent periodic solutions of the relay servomechanism considered. It is, however, equally apparent that these periodic solutions are unstable: the slightest disturbance will cause the lines of behavior of the system to spiral to the origin or to

diverge to infinity. Therefore such periodic solutions cannot occur in reality.

One property of our solution for the optimum switching problem stands out: for all cases, the optimum switching function φ takes the value -1 in the first quadrant of the phase plane and takes the value $+1$ in the third quadrant. By writing Eq. (10.19) in the following form,

Fig. 10.32

$$\frac{d^2y}{dt^2} = -2\zeta\frac{dy}{dt} - y + \varphi\left(y, \frac{dy}{dt}\right)$$

we can understand this feature of the solution quite easily. The purpose of the design is to return to the state $y = 0$, or the t axis, in the shortest possible time. When y and dy/dt are both positive, this purpose can be accomplished by making d^2y/dt^2, or the curvature of the $y(t)$ curve, as negatively large as possible, that is, φ should be -1. When y and dy/dt are both negative, d^2y/dt^2 should be as positively large as possible, that is, φ should be $+1$. This intuitive reasoning agrees with our result for the optimum switching function. When y and $\dot{y}$ have different signs, the optimum value of φ cannot be so simply determined, because then the rate of return to the t axis depends upon the complicated interaction between y and $\dot{y}$. Bushaw's contribution is then to specify the optimum

φ just in these regions, the second and the fourth quadrants of the phase plane. But it is clear from this discussion that the optimum switching line C must lie in the second and the fourth quadrants.

For systems of higher order and for systems of more than one degree of freedom, the phase-plane representation of the state of the system is no longer possible. We have to use a phase space of many dimensions. By analogy with the problem already discussed, we may expect the solution of optimum switching of such complicated systems to be the determination of optimum switching surfaces in the phase space. However, such problems have not yet been solved; only an initial attempt has been made by Kang and Fett.[1]

10.9 Multiple-mode Operation. What happens when the system is guided by our switching action to the origin of the phase plane? Clearly,

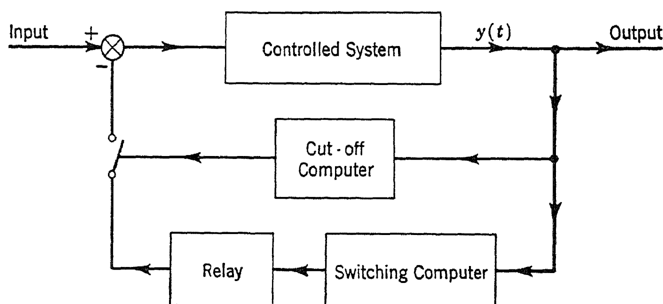


FIG. 10.33

if the forcing function is continued at the value obtaining just before the origin is reached, the system will move away from the desired state again. But as soon as it moves away from the origin, the switching action designed into the system will function and drive the system back towards the origin. The net result will be a rapid approach to the origin from any disturbed position, and then a high-frequency oscillation or chattering around the origin.

If small deviations from the rest state or the origin are not objectionable, the chattering near the origin can be eliminated by cutting off the forcing function whenever the system is near the rest state, *i.e.*, whenever y and $\dot{y}$ are small enough to be negligible. The system then has two modes of operation: for large deviations, the switching action together with the output of the relay functions; for small deviations, such input to the system is cut off. We shall see the necessity of such *multiple-mode operation* of servomechanisms again in the later chapters.

The optimum switching line, being nonlinear, cannot be implemented by a simple linear circuit. In fact, the measured value of the output $y(t)$

[1] C. L. Kang and G. H. Fett, *J. Appl. Phys.*, **24**, 38–41 (1953).

has to be "digested" by a nonlinear device, or a *computer*. The computer is designed to generate the switching signal for the relay according to the optimum switching line. Furthermore, there should be another *cutoff computer* to disconnect the relay output from the system when the output y and $\dot{y}$ are small enough. Therefore the block diagram of such a relay servomechanism is that of Fig. 10.33. The incorporation of computers into control systems will be generally necessary for the more complicated systems discussed in the following chapters. However, conceptually, this really involves nothing new—the compensating circuit used in a conventional servomechanism for modifying the system transfer function is also a computer. But in these simpler systems, the computing function can be performed by a linear circuit, such as an RC circuit. We shall give a more extensive discussion on computers in Chap. 13.

CHAPTER 11

NONLINEAR SYSTEMS

A nonlinear system is a system for which the output is not linearly proportional to the input. The relay servomechanism is a simple example of such nonlinear systems. In Chap. 6, we have given a general method of linearizing any nonlinear servomechanism, *i.e.*, any nonlinear system can be made to behave like a linear system by modifying the system into an oscillating-control servomechanism. In the preceding chapter, we have presented a method for analyzing a servomechanism including a nonlinear device whose behavior is insensitive to the frequency of the input. These methods for designing nonlinear servomechanisms encompass a wide variety of nonlinear problems in engineering practice and are quite sufficient for dealing with the usual systems synthesis problems.

On the other hand, as indicated in the later sections of the last chapter, the problem of optimum utilization of the nonlinear characteristics of a system to improve the performance of the system is generally a much more difficult problem than the problem of design for stability. In fact, only a modest beginning has been made in this direction. It is thus not possible at the present time to give a satisfactory treatment of the subject of general nonlinear servomechanisms. Furthermore, the problem should perhaps be formulated more directly and in a different way. Instead of finding the performance of an assumed system, we should specify the performance of the system, and then determine the required nonlinearity. This approach will be discussed in Chap. 14. The scope of this chapter is quite limited: we shall indicate only a few possibilities of purposefully utilizing the characteristics of a nonlinear system.

11.1 Nonlinear Feedback Relay Servomechanism. If we confine ourselves to cases where the deviations from the equilibrium state are small, the results of Bushaw stated in Sec. 10.8 for the optimum switching line of a relay servomechanism can be greatly simplified. It is seen from the discussions in that section that near the origin the switching line C is approximated by

$$\dot{y}|\dot{y}| = -2y \tag{11.1}$$

This indicates that an improvement in the performance of the system over that of a system with linear switching (Sec. 10.6) can be achieved

by using nonlinear switching defined by Eq. (11.1). If x is the input and y the output, Eq. (11.1) should be modified into

$$a^2\dot{y}|\dot{y}| = (x - y) \qquad (11.2)$$

or

$$\operatorname{sgn}(x - y)\sqrt{|x - y|} = a\dot{y} \qquad (11.3)$$

where a is a constant.

The block diagram of a relay servomechanism utilizing the switching condition of Eq. (11.2) is shown in Fig. 11.1. This method of $\dot{y}|\dot{y}|$ feedback was proposed by Uttley and Hammond[1] for improving the performance of the simple relay servomechanism. The computer here

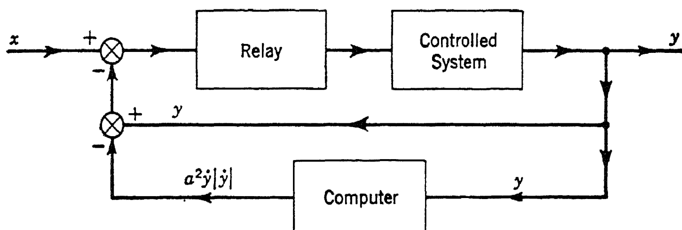


FIG. 11.1

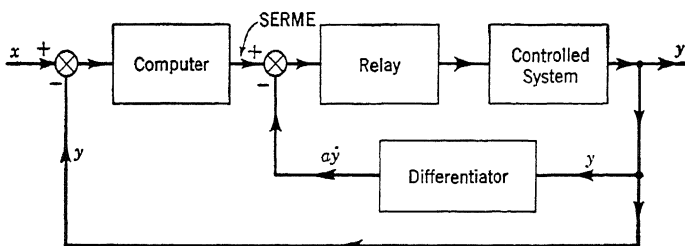


FIG. 11.2

is only required to generate the signal $a^2\dot{y}|\dot{y}|$ and thus can be relatively simple. The equivalent switching condition of Eq. (11.3) can also be implemented by the $\dot{y}$ and $\operatorname{sgn}(x - y)\sqrt{|x - y|}$ feedback. Since $(x - y)$ is the error, the system may be called *sign error root-modulus error* (*SERME*) system. The block diagram is shown in Fig. 11.2. Here again, the computer is relatively simple. This system was proposed by J. C. West.[2]

These nonlinear feedback relay servomechanisms, although comparatively simple, cannot be analyzed rationally. In fact, our argument in their favor is only a plausible one, not a complete one. The final design

[1] A. M. Uttley, P. H. Hammond, "Automatic and Manual Control," p. 285, edited by A. Tustin, Butterworth & Co. (Publishers) Ltd., London, 1952.

[2] J. C. West, "Automatic and Manual Control," p. 300, edited by A. Tustin, Butterworth & Co. (Publishers) Ltd., London, 1952.

of any particular system using these principles has to be determined by actual experimentation.

11.2 Systems with Small Nonlinearity. If a system of n degrees of freedom is not linear, then instead of a linear differential equation, such as Eq. (2.3), we have

$$a_n \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_1 \frac{dy}{dt} + a_0 y + \mu f\left(y, \frac{dy}{dt}, \ldots, \frac{d^{n-1}y}{dt^{n-1}}\right) = x(t) \quad (11.4)$$

where the a's and μ are constants, $x(t)$ is the input, $y(t)$ is the output, and f is a nonlinear function of its variables. The first part on the left of Eq. (11.4) is thus the linear differential operator, as in Eq. (2.3). All the nonlinearity of the system is represented by the last term on the left of the equation. Small nonlinearity means that μ is small in comparison with the a's.

For small nonlinearity then, we may try a formal expansion of the solution in a power series of μ:

$$y(t) = y^{(0)}(t) + \mu y^{(1)}(t) + \mu^2 y^{(2)}(t) + \cdots \quad (11.5)$$

By substituting Eq. (11.5) into Eq. (11.4) and by equating equal powers of μ, we have

$$a_n \frac{d^n y^{(0)}}{dt^n} + a_{n-1} \frac{d^{n-1}y^{(0)}}{dt^{n-1}} + \cdots + a_1 \frac{dy^{(0)}}{dt} + a_0 y^{(0)} = x(t) \quad (11.6)$$

$$a_n \frac{d^n y^{(1)}}{dt^n} + a_{n-1} \frac{d^{n-1}y^{(1)}}{dt^{n-1}} + \cdots + a_1 \frac{dy^{(1)}}{dt} + a_0 y^{(1)} = -f\left(y^{(0)}, \ldots, \frac{d^{n-1}y^{(0)}}{dt^{n-1}}\right) \quad (11.7)$$

and other equations for higher-order terms. The "zeroth-order" approximation is thus the linear equation, as in Eq. (2.3). But more important is the fact that the first-order correction term due to nonlinearity is determined by Eq. (11.7), an equation of exactly the same characteristics as the zeroth-order approximation. In other words, if the linear approximation shows that the system is damped and has other desired properties of a servo system, then the first-order correction $y^{(1)}(t)$ will also have such properties. Moreover, because of the occurrence of the small parameter μ before $y^{(1)}(t)$ in the expansion of Eq. (11.5), the corrections for nonlinear effects are small. Consequently, small nonlinearity in a system of "satisfactory" performance will not alter essentially the system behavior from its linear approximation. Therefore, as far as an engineering approximation is concerned, we can treat such systems as linear systems. This is the reason that the linear theory of servomech-

anisms has had such good success in spite of the ever present small nonlinearity in even a "linear" system.

On the other hand, if the damping of the linear approximation is very small, then we know that there is the possibility of resonance. That is, even if the input $x(t)$ is of order unity, the output $y^{(0)}(t)$ of the linear approximate system of Eq. (11.6) can be very much larger than unity. When this is the case, the quantity $\mu f\left(y^{(0)}, \ldots, \frac{d^{n-1}y^{(0)}}{dt^{n-1}}\right)$, or the nonlinear effects, can be of the same order of magnitude as some of the linear terms, even with small μ. In other words, our formal expansion procedure of the last paragraph is not justified, and we must expect strong

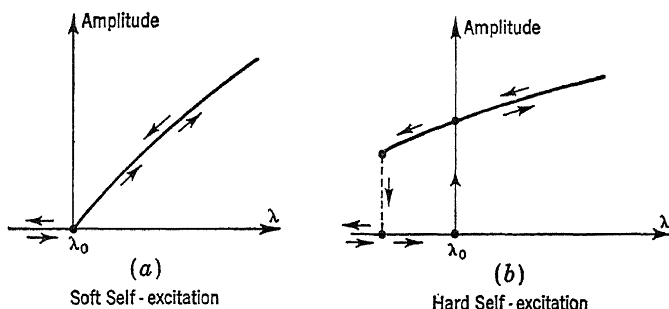


Fig. 11.3

effects from even small nonlinearity if the system has only very weak damping. In the following sections, we shall give a brief description of the kaleidoscopic behavior of a nonlinear system. Detailed treatment of such phenomena in nonlinear mechanics can be found in the excellent books by Minorsky[1] and Stoker.[2]

11.3 Jump Phenomenon. As already stated in Sec. 10.2, self-excited oscillation in a system, if it is automatically built up from very small deviations from the equilibrium state, is called "soft" self-excited oscillation; if it has to be started by large deviations from the equilibrium state, the oscillation is called "hard" self-excited oscillation. In some cases, the coefficients of the differential equation of the system may depend upon a parameter λ of the system. If, for some particular value of λ, the critical λ_0, the character of the equilibrium state changes from that of a stable state to an unstable state, then the limit cycle or steady oscillation will appear. According to whether the system is one of soft or hard self-excitation, the phenomenon occurs as in either Fig. 11.3*a* or *b*. In the

[1] N. Minorsky, "Introduction to Nonlinear Mechanics," Edwards Bros., Inc., Ann Arbor, Mich., 1947.

[2] J. J. Stoker, "Nonlinear Vibrations," Interscience Publishers, Inc., New York, 1950.

first case, if the parameter λ increases, nothing happens until λ reaches the value λ_0, at which point the equilibrium state changes from stability to instability with a simultaneous appearance of a stable limit cycle whose amplitude begins to increase with λ. If λ is made to decrease, the phenomenon retraces its path exactly, and the limit cycle disappears when $\lambda = \lambda_0$. For a system with hard self-excitation, the picture is different (Fig. 11.3*b*). The oscillation appears suddenly at $\lambda = \lambda_0$ with a finite amplitude, and, with increasing λ, the amplitude increases. If λ decreases, the oscillation or limit cycle does not disappear when $\lambda = \lambda_0$ but disappears farther along the curve where the amplitude again jumps from a finite value to zero. Thus the jump phenomenon is associated with the hysteresis of the system behavior.

11.4 Frequency Demultiplication. If a nonlinear system is acted on by a periodic input containing two frequencies ω_1 and ω_2, the output of the system occurs not only with these frequencies and their harmonics but also with an additional spectrum of the so-called *combination tones* $m\omega_1 \pm n\omega_2$, where m and n are integers. Thus, for example, if a voltage $x = x_0(\cos \omega_1 t + \cos \omega_2 t)$ is impressed on a nonlinear conductor whose current y is given by $y = a_1x + a_2x^2 + a_3x^3$, then the output $y(t)$ will contain the following frequencies: ω_1, ω_2, $2\omega_1$, $2\omega_2$, $3\omega_1$, $3\omega_2$, $\omega_1 + \omega_2$, $\omega_1 - \omega_2$, $2\omega_1 + \omega_2$, $2\omega_1 - \omega_2$, $\omega_1 + 2\omega_2$, and $\omega_1 - 2\omega_2$. The first six frequencies are regular harmonics, but the last six are the combination tones resulting from the nonlinearity of the conductor. Some of these are higher, and others are lower than the original frequencies ω_1 and ω_2. These lower frequencies are called *subharmonics*, and the process by which they are obtained is called *frequency demultiplication.*

It is seen that if ω_1 and ω_2 are fairly close together, $\omega_1 - \omega_2$ can be much smaller than either of the original frequencies. If, in addition, the system can be stabilized at such a low frequency, subharmonics of the order of $\frac{1}{100}$ of the input frequency, and even lower, can be obtained. If several such systems are connected in series, with the subharmonic output of one serving as input to the other, then still lower frequencies can be reached.

11.5 Entrainment of Frequency. If a nonlinear system has a self-excited oscillation of frequency ω_1, then when the system is subjected to an input of a slightly different frequency ω_2, we expect the simultaneous occurrence of both ω_1 and ω_2, and, through nonlinear interaction, the beat frequency $\omega_2 - \omega_1$. In reality the phenomenon develops as shown in Fig. 11.4. The frequency $\omega_2 - \omega_1$ disappears suddenly as soon as ω_2 reaches a certain *zone of synchronization AB*. In this interval there exists only one frequency ω_2, and everything happens as if the original frequency ω_1 were entrained by the variable frequency ω_2.

This phenomenon of *frequency entrainment* was first explained by van

der Pol and was extended by others. Let the system be one of second order and determined by the differential equation

$$\frac{d^2y}{dt^2} - \alpha\frac{dy}{dt} + \gamma\left(\frac{dy}{dt}\right)^3 + \omega_1^2 y = B\omega_1^2 \sin \omega_2 t \tag{11.8}$$

where α, γ, and B are positive constants. If $B = 0$, the system has negative damping for small amplitudes of oscillation, but positive damping for large amplitudes of oscillation. Thus there is an amplitude at which the system can maintain a steady oscillation. Furthermore, if both α and γ are small, then steady self-excited oscillation must occur at a frequency close to ω_1. Van der Pol showed that when ω_2 is close to ω_1, the solution of Eq. (11.8) can be written as

$$y(t) = b_1(t) \sin \omega_2 t + b_2(t) \cos \omega_2 t \tag{11.9}$$

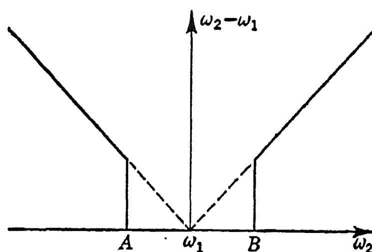


FIG. 11.4

and that $b_1(t)$ and $b_2(t)$ are slowly varying functions of t, such that, for example,

$$\left|\frac{db_1}{dt}\right| \ll \left|\omega_2 b_1(t)\right|$$

By substituting Eq. (11.9) into Eq. (11.8) and retaining only the terms up to first order, we can write the equations for b_1 and b_2 as

$$\left.\begin{aligned}\frac{db_1}{dt} &= f_1(b_1,b_2;\omega_2)\\ \frac{db_2}{dt} &= f_2(b_1,b_2;\omega_2)\end{aligned}\right\} \tag{11.10}$$

These are autonomous equations of the first order. Therefore they can be solved by the method of isoclines in the same manner as a second-order system in the phase plane. Analysis shows that for a certain range of ω_2 near ω_1, the system of Eq. (11.10) has a stable node point in the b_1b_2 plane. Thus, no matter what the initial values of b_1 and b_2 are, the system tends to a fixed set of values of b_1 and b_2 corresponding to this point. Thus no oscillations of the frequency $\omega_2 - \omega_1$ appear at all. For ω_2 outside this range, there is a stable limit cycle in the b_1b_2 plane which accounts for the observed effect shown in Fig 11.4.

11.6 Asynchronous Excitation and Quenching. In some nonlinear systems it is possible either to start or to stop an oscillation of frequency ω by means of another oscillation of an entirely different frequency ω_1.

In the first case, the phenomenon is called *asynchronous excitation*, and in the second, *asynchronous quenching*. An understanding of these phenomena can be obtained by recalling the fact that the mere existence of a limit cycle in the phase space of a system does not necessarily mean the occurrence of steady oscillation. For steady oscillations to take place, the limit cycle must be stable in the sense that the system has a tendency to return to the limit cycle, even if the system is disturbed and displaced to a point away from the limit cycle in the phase space. An unstable limit cycle cannot be realized in a physical system. In view of this, one can readily see that the appearance of a new oscillation may, under certain circumstances, either create or destroy the conditions of stability of another. In the first case we have the phenomenon of asynchronous excitation and in the second that of asynchronous quenching. The term asynchronous merely emphasizes the fact that the relation between ω and ω_1 is purely arbitrary.

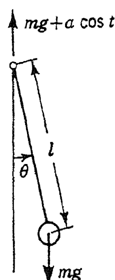


FIG. 11.5

11.7 Parametric Excitation and Damping. It has been known for a long time that if a parameter of an oscillating system is varied periodically with a frequency ω, the system begins to oscillate at the frequency $\omega/2$. Lord Rayleigh demonstrated this effect by attaching one end of a stretched wire to a prong of a tuning fork. If the fork oscillates with frequency ω, lateral oscillations of the wire appear with frequency $\omega/2$. A similar problem is that of a simple pendulum, a mass supported by a weightless bar, under the influence of a sinusoidal force applied at the end of the bar (Fig. 11.5). If θ is the small angular displacement of the pendulum from the vertical, and, without loss of generality, if the frequency of the sinusoidal force is unity, the differential equation for θ is then

$$ml\frac{d^2\theta}{dt^2} + (mg + a\cos t)\theta = 0$$

where m is the mass, g the gravitational attraction, l the length of the pendulum, and a the amplitude of the periodic force. This equation can be written as

$$\frac{d^2\theta}{dt^2} + (\alpha + \beta\cos t)\theta = 0 \tag{11.11}$$

α is thus equal to g/l and β is equal to a/ml. The case of an inverted pendulum, with the mass above the point of support, can also be represented by Eq. (11.11) by taking g as negative. Therefore, for a normal pendulum α is positive, and for an inverted pendulum α is negative. Equation (11.11) is actually a linear equation but with periodic variation in the end force applied to the pendulum. Thus the system can be considered as one with periodic variation of its parameter.

Equation (11.11) is the well-known Mathieu differential equation. The stability of the solution is determined by the constants α and β. In fact, the $\alpha\beta$ plane can be divided into a region of stability and a region of instability as shown in Fig. 11.6 (where the stable region is shaded). It is thus seen that for positive α, the case of the normal pendulum, the system is stable when the periodic end force is absent, or $\beta = 0$. This is, of course, well known. However, it is interesting to note that with appropriate β, the system can be made unstable. Then the pendulum will swing with increasing amplitude until, finally, nonlinear effects limit

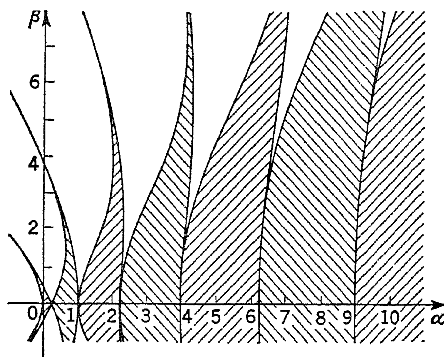


FIG. 11.6

the amplitude at a fixed large value. This is the phenomenon of *parametric excitation.* For negative values of α, the case of inverted pendulum, the system without a periodic end force is naturally unstable. But for a certain narrow range of β, the system can be stabilized. This phenomenon is called *parametric damping.*

Parametric excitation or parametric damping can occur in any system with periodic variation in the parameter of the system. This phenomenon or any of the previously shown nonlinear phenomena can be utilized to achieve the desired performance of a control system. In fact, many of them are already incorporated into the many *components* of a servo control system. However, these nonlinear components are mere "gadgets," and they are "designed" rather by experience and testing than by theoretical analysis. The application of the characteristically nonlinear phenomena to the over-all design of a control system is as yet an unexplored field. Our discussion in the preceding sections serves only as an indication of its rich possibilities.

CHAPTER 12

LINEAR SYSTEM WITH VARIABLE COEFFICIENTS

The only system with time-varying coefficients considered in detail in the previous chapters is the pendulum with a periodic force at the supporting end, discussed in connection with the phenomena of parametric excitation and damping. All other systems considered do not have coefficients of their differential equations that are explicitly functions of time. We have, however, shown in Chap. 1 that linear systems with time-varying coefficients can have behavior entirely different from that of systems with constant coefficients. In this chapter, we shall again take up this question and discuss in some detail such a typical but simple system: the short-range artillery rocket. We shall demonstrate that the question of stability of such a system with variable coefficients cannot be solved in the same manner as for the linear system with constant coefficients. Not only is the method of Laplace transform and transfer function useless for the purpose, but we are forced to change our entire approach to the problem.

We shall study the motion of a fin-stabilized artillery rocket during the period of action of the rocket thrust. We shall be particularly concerned about the angular deviations of the rocket axis from the launching angle caused by disturbances during the launching and the subsequent damping action of the fins. The general problem of the dynamics of artillery rockets was studied in great detail by various authors in different countries during World War II. The American work is summarized by Rosser, Newton, and Gross.[1] The work done in England is reported by Rankin.[2] Carrière's paper[3] represents a French investigation of the same subject. Our discussion here will be greatly simplified and has the purpose of bringing out only the salient points of interest to our study of linear systems with variable coefficients.

12.1 Artillery Rocket during Burning. For a fin-stabilized artillery rocket, the interaction of motion in the vertical plane and the horizontal plane is negligible, *i.e.*, there are negligible aerodynamic forces in the

[1] J. B. Rosser, R. R. Newton, and G. L. Gross, "Mathematical Theory of Rocket Flight," McGraw-Hill Book Company, Inc., New York, 1947.

[2] R. A. Rankin, *Trans. Roy. Soc. London (A)*, **241**, 457–585, (1949).

[3] P. Carrière, *Mém. artillerie franç.*, **25**, 253–360 (1951).

vertical plane produced by the motion in the horizontal plane. Therefore the characteristics of the rocket are not lost by confining our considerations to the vertical plane and studying the motion in that plane only. We shall consider the earth to be flat for short-range artillery rockets. Let v be the magnitude of the velocity of the rocket, θ the inclination of the velocity vector, and ϕ the inclination of the axis of the rocket (Fig. 12.1). Then the angle of attack α of the rocket is

$$\alpha = \phi - \theta \tag{12.1}$$

Let m be the mass of the rocket, and g the gravitational constant. The gravitational attraction is thus a vertically downward force mg. The

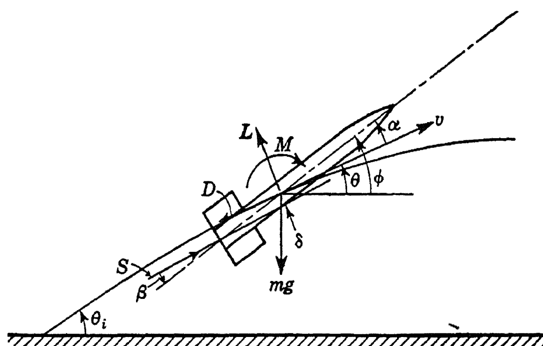


FIG. 12.1

aerodynamic forces are the lift L, the drag D, perpendicular and parallel to the direction of motion, respectively, and the moment M. All these forces act on the center of gravity of the rocket. The constant thrust S of the rocket motor may have an angular misalignment β and a moment arm δ with respect to the center of gravity, as indicated in Fig. 12.1. Then the equation for acceleration along the trajectory is

$$m \frac{dv}{dt} = S \cos (\alpha - \beta) - mg \sin \theta - D \tag{12.2}$$

The equation for acceleration normal to the trajectory is

$$mv \frac{d\theta}{dt} = S \sin (\alpha - \beta) - mg \cos \theta + L \tag{12.3}$$

Finally, if k is the radius of gyration of the rocket about a lateral axis through the center of gravity, the equation of angular acceleration is

$$mk^2 \frac{d^2\phi}{dt^2} = S\delta - M \tag{12.4}$$

In Eq. (12.4), we have neglected the so-called jet damping moment since it is small in comparison to the restoring moment of the tail fins.

The aerodynamic forces and the moment are dependent upon the angle of attack α. But if the fins on the rocket are not properly aligned, the lift and moment will not vanish even if α is zero. The misalignment can be accounted for by introducing a misalignment angle γ, such that L and M vanish not at $\alpha = 0$ but at $\alpha = \gamma$. If ρ is the density of air and d the diameter of the rocket body, we can introduce the lift coefficient K_L, the drag coefficient K_D, and the moment coefficient K_M as follows:

$$L = K_L \rho v^2 d^2 \sin(\alpha - \gamma) \tag{12.5}$$

$$D = K_D \rho v^2 d^2 \tag{12.6}$$

$$M = K_M \rho v^2 d^3 \sin(\alpha - \gamma) \tag{12.7}$$

For short-range artillery rockets, the summit of the trajectory is not high; therefore the density ρ can be taken as a constant. Furthermore, the maximum velocity is small enough so that all the coefficients K_L, K_D, and K_M can be considered as constants, not influenced by the Mach number on the trajectory. Moreover, for short-range rockets, the propellant fraction of the total mass is small; then the mass m of the rocket can be taken to be a constant without serious error. Equations (12.1) to (12.7), together with these simplifying assumptions, then determine the trajectory of the rocket.

The burning time of this type of rockets is very short, say $\frac{1}{5}$ of a second. This necessarily makes the acceleration S/m very large. In fact, the thrust S is so large that the gravitational and drag forces are negligibly small in comparison. Furthermore, the deviation of thrust line from the flight direction, $\alpha - \beta$, is never large. Therefore an immediate zeroth-order approximation of Eq. (12.2) is

$$\frac{dv}{dt} = \frac{S}{m} \tag{12.8}$$

Therefore the zeroth-order approximation to the trajectory is given by a straight line with the initial inclination θ_i (Fig. 12.1). The motion along this straight line is one of constant acceleration S/m. Thus if z is the distance measured along this line, then the motion is represented by

$$v^2 = \frac{2S}{m} z = \left(\frac{S}{m}\right)^2 t^2 \tag{12.9}$$

If the rocket is launched without any initial velocity, then z is the true distance from the launching point. If there is initial velocity, z is not the distance measured from the launching point, but from some point ahead of it. From Eq. (12.9) we have

$$\frac{d}{dt} = \frac{dz}{dt}\frac{d}{dz} = v\frac{d}{dz} = \sqrt{\frac{2S}{m} z}\,\frac{d}{dz} \tag{12.10}$$

With this zeroth-order solution, we can calculate the first-order solution of the rocket trajectory. This will be done presently.

12.2 Linearized Trajectory Equations. Since the deviation of the trajectory from the zeroth-order solution during burning is never large, we can replace the velocity v and the time derivative whenever they occur in Eqs. (12.3) and (12.4) by those given by Eqs. (12.9) and (12.10). Furthermore, since $\alpha - \beta$ is small, $\sin(\alpha - \beta)$ can be replaced by $\alpha - \beta$. $\cos\theta$ can be replaced by $\cos\theta_i$. We shall also neglect the lift L, since it is small in comparison to the lateral thrust component and the weight of the rocket. With these simplifications, Eqs. (12.3) and (12.4) become

$$2z\frac{d\theta}{dz} = (\alpha - \beta) - \frac{mg}{S}\cos\theta_i \tag{12.11}$$

and

$$2z\frac{d^2\phi}{dz^2} + \frac{d\phi}{dz} = \frac{\delta}{k^2} - \frac{8\pi^2}{\sigma^2}z(\alpha - \gamma) \tag{12.12}$$

where σ is defined by

$$\sigma^2 = 4\pi^2\frac{k^2m}{K_M\rho d^3} \tag{12.13}$$

and evidently has the dimension of a length. σ can be taken as the characteristic length of the disturbed motion of the rocket, and may be considered as the wave length of the disturbed trajectory. Equations (12.1), (12.11), and (12.12) are then linearized equations for the three unknowns α, θ, and ϕ. Linearization is made under the assumption of small deviations from the straight-line trajectory at the ideal launching angle θ_i.

We can eliminate θ and ϕ to obtain a single equation for α. To do this, we divide Eq. (12.11) by $2\sqrt{z}$ and differentiate the resultant equation with respect to z. Then we have

$$\sqrt{z}\frac{d^2\theta}{dz^2} + \frac{1}{2\sqrt{z}}\frac{d\theta}{dz} = \frac{1}{2\sqrt{z}}\frac{d\alpha}{dz} - \frac{1}{4z\sqrt{z}}\left(\alpha - \beta - \frac{gm}{S}\cos\theta_i\right)$$

Now we divide Eq. (12.12) by $2\sqrt{z}$ and subtract from it the above equation. Then, using Eq. (12.1), we have

$$\sqrt{z}\frac{d^2\alpha}{dz^2} + \frac{1}{\sqrt{z}}\frac{d\alpha}{dz} + \left(\frac{4\pi^2\sqrt{z}}{\sigma^2} - \frac{1}{4z\sqrt{z}}\right)\alpha = \frac{\delta}{2k^2\sqrt{z}} + \frac{4\pi^2\sqrt{z}}{\sigma^2}\gamma - \frac{1}{4z\sqrt{z}}\left(\beta + \frac{gm}{S}\cos\theta_i\right)$$

This equation clearly demonstrates the fact that the controlling differential equation of the stability of an artillery rocket is not an equation of constant coefficients. In fact, this equation can be put into the

standard form of Bessel's equation by introducing the nondimensional distance ξ, defined by

$$\xi = \frac{2\pi z}{\sigma} \tag{12.14}$$

where σ is the "wave length" specified by Eq. (12.13). Then we have

$$\frac{d^2\alpha}{d\xi^2} + \frac{1}{\xi}\frac{d\alpha}{d\xi} + \left[1 - \frac{(1/2)^2}{\xi^2}\right]\alpha = \gamma + \left(\frac{\delta\sigma}{4\pi k^2}\right)\frac{1}{\xi} - \frac{1}{4}\left(\beta + \frac{gm}{S}\cos\theta_i\right)\frac{1}{\xi^2} \tag{12.15}$$

When α is determined, θ can be calculated by integrating the following equation, obtained from Eq. (12.11):

$$\frac{d\theta}{d\xi} = \frac{1}{2\xi}\left[\alpha - \beta - \frac{mg}{S}\cos\theta_i\right] \tag{12.16}$$

The independent variable z or ξ in the differential equations is not a time variable but a distance variable. However, since, as shown by Eq. (12.9), z and thus ξ are monotonic functions of t, the stability of the system is not modified by changing the independent variable from t to ξ; if the system is stable in terms of ξ, it is stable in terms of t, in the sense that the deviations from the ideal straight-line trajectory will decrease with increase in ξ or t. Therefore the problem of stability can be adequately discussed with the variable ξ. ξ increases from the initial value ξ_0 at $t = 0$ as time increases. The initial value ξ_0 is zero if the launching velocity is zero.

12.3 Stability of an Artillery Rocket. To discuss the question of stability, we must solve Eqs. (12.15) and (12.16) with the specified initial conditions and then determine whether the angle of attack α or, more appropriate, the deviation of the inclination of the trajectory, $\theta - \theta_i$, tends to zero as ξ increases. Equation (12.15) is actually Bessel's equation for the order $\frac{1}{2}$. The complementary functions are thus Bessel functions of order $\frac{1}{2}$ and $-\frac{1}{2}$. They are, however, expressible in terms of elementary functions. In fact, Eq. (12.15) can be rewritten as

$$\frac{d^2\zeta}{d\xi^2} + \zeta = Q(\xi) \tag{12.17}$$

where

$$\zeta = \sqrt{\xi}\,\alpha \tag{12.18}$$

and

$$Q(\xi) = \gamma\sqrt{\xi} + \left(\frac{\delta\sigma}{4\pi k^2}\right)\frac{1}{\sqrt{\xi}} - \frac{1}{4}\left(\beta + \frac{gm}{S}\cos\theta_i\right)\frac{1}{\xi^{\frac{3}{2}}} \tag{12.19}$$

Therefore the complementary functions for ζ are simply $\sin\xi$ and $\cos\xi$.

The conditions of the rocket as it leaves the launcher, or the initial conditions, are

$$\left.\begin{aligned} v &= v_0 \\ \theta &= \theta_0 \\ \alpha &= \alpha_0 \\ \text{and} \qquad d\phi/dt &= (d\phi/dt)_0 \end{aligned}\right\} \tag{12.20}$$

The subscript $_0$ thus indicates quantities at the instant $t = 0$. The initial values of ξ and ζ are thus

$$\xi_0 = \frac{2\pi}{\sigma}\frac{m}{2S}v_0^2 = \frac{\pi m v_0^2}{\sigma S} \tag{12.21}$$

and

$$\zeta_0 = \sqrt{\xi_0}\,\alpha_0 \tag{12.22}$$

By using Eq. (12.16), we have at $t = 0$

$$\sqrt{\xi_0}\left(\frac{d\theta}{d\xi}\right)_0 = \frac{\alpha_0 - \beta - (mg/S)\cos\theta_i}{2\sqrt{\xi_0}}$$

But $\theta = \phi - \alpha$, thus

$$\sqrt{\xi_0}\left(\frac{d\alpha}{d\xi}\right)_0 + \frac{1}{2\sqrt{\xi_0}}\alpha_0 = \left(\frac{d\zeta}{d\xi}\right)_0 = \sqrt{\xi_0}\left(\frac{d\phi}{d\xi}\right)_0 + \frac{\beta + (gm/S)\cos\theta_i}{2\sqrt{\xi_0}}$$

or explicitly, then

$$\left(\frac{d\zeta}{d\xi}\right)_0 = \sqrt{\xi_0}\,\frac{\sigma(d\phi/dt)_0}{2\pi v_0} + \frac{\beta + (gm/S)\cos\theta_i}{2\sqrt{\xi_0}} \tag{12.23}$$

With the initial conditions so translated, we can write down the solution of Eq. (12.17) or ζ or α immediately,

$$\begin{aligned} \alpha(\xi,\xi_0) &= \frac{1}{\sqrt{\xi}}\cos(\xi - \xi_0)\left[\zeta_0 - \int_{\xi_0}^{\xi}\sin(\eta - \xi_0)Q(\eta)\,d\eta\right] \\ &\quad + \frac{1}{\sqrt{\xi}}\sin(\xi - \xi_0)\left[\left(\frac{d\zeta}{d\xi}\right)_0 + \int_{\xi_0}^{\xi}\cos(\eta - \xi_0)Q(\eta)\,d\eta\right] \end{aligned} \tag{12.24}$$

where Q is the forcing function specified by Eq. (12.19). Since $Q(\eta)$ contains half powers of η, the integrals of Eq. (12.24) are actually Fresnel integrals. When α is calculated, Eq. (12.16) then gives θ by quadrature,

$$\theta - \theta_i = (\theta_0 - \theta_i) - \frac{1}{2}\left(\beta + \frac{mg}{S}\cos\theta_i\right)\log\frac{\xi}{\xi_0} + \frac{1}{2}\int_{\xi_0}^{\xi}\frac{\alpha(\eta,\xi_0)}{\eta}\,d\eta \tag{12.25}$$

We can further separate the effects of different disturbances at the launcher by writing Eq. (12.25) in a series of terms, each representing

one type of disturbance. Thus

$$\theta - \theta_i = (\theta_0 - \theta_i) + \left(\beta + \frac{mg}{S}\cos\theta_i\right) G_1(\xi,\xi_0) + \frac{\delta\sigma}{2\pi k^2} G_2(\xi,\xi_0)$$
$$- \gamma[G_1(\xi,\xi_0) + G_3(\xi,\xi_0)] + \alpha_0 G_3(\xi,\xi_0) + \frac{1}{2}\frac{\sigma}{\pi v_0}\left(\frac{d\phi}{dt}\right)_0 G_4(\xi,\xi_0) \quad (12.26)$$

The first term represents the effect of the initial deviation of the trajectory angle. The second term gives the effect of thrust misalignment and

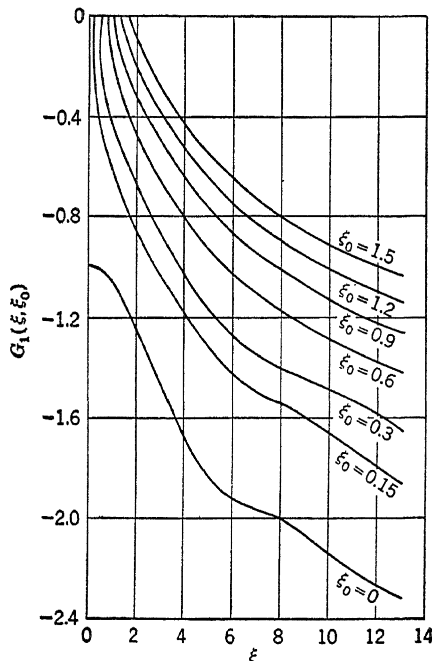


FIG. 12.2

gravitational pull. The third term shows the influence of the turning moment of the thrust. The fourth term gives the effect of fin misalignment. The fifth term represents the effect of the initial angle of attack. The last term gives the influence of the initial angular velocity of the rocket. Each of the G's is a function of the two variables ξ and ξ_0, and is a combination of Fresnel's integrals. Rosser and his collaborators call them the "rocket functions" and tabulate them in their book. Figures 12.2 to 12.5, representing these functions graphically, are also taken from this book.

An inspection of Figs. 12.2 to 12.5 immediately brings out the fact that all the G functions have persistent values for large ξ. Thus the disturbances do not damp out. The first and the last two terms of

Eq. (12.26) represent the effects of the initial disturbances at the launching point. They have finite nonvanishing values at large ξ. The other terms in Eq. (12.26) are the "output" due to "input," or the forcing functions. They also have values for large ξ. The behavior of the G_1 function is even worse: at large ξ, it is approximately log ξ, and thus increases without limit. Therefore, if we use the previously established

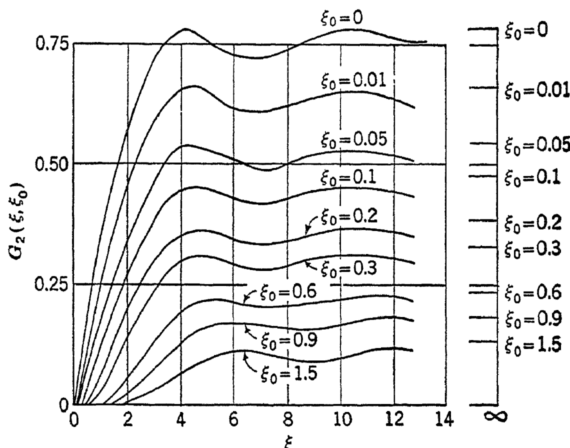


FIG. 12.3

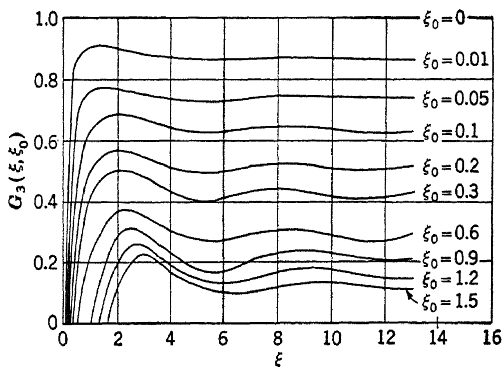


FIG. 12.4

criterion for stability of a system, *i.e.*, the vanishing of initial disturbances and the boundedness of the output under "reasonable" forcing functions, then the artillery rocket is unstable. On the other hand, the complementary functions of the basic equation, Eq. (12.15), are Bessel functions which vanish for large values of the variable and thus seem to indicate stability of the system. If we try to apply our experience with systems which are described by a linear differential equation with constant coefficients, the behavior of a system with variable coefficients

apparently is confusing. However, this only indicates the inapplicability of concepts developed for systems with constant coefficients to systems with time-varying coefficients. A new approach to the problem of stability and control is required. A discussion on this point will be presented in the following section.

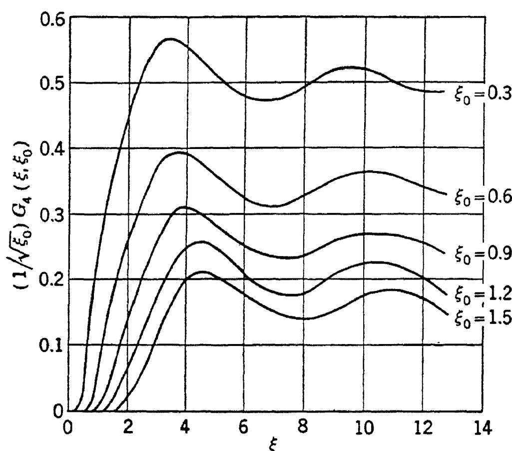


FIG. 12.5

12.4 Stability and Control of Systems with Variable Coefficients. For a linear system with constant coefficients, our previous study has shown that the satisfactory performance of the system with respect to stability and other control criteria can be guaranteed if the solutions of the homogeneous equation without forcing function or input are all sufficiently damped. Therefore, although the forcing function or input may vary considerably from one case to another, the performance criteria and thus the design of the system are based upon the study of the solutions of the homogeneous equation or the complementary functions of the complete equation. This is the fundamental principle in the theory of servomechanisms. The actual method of analysis by the technique of transfer functions based upon the Laplace transform is merely a useful trick. In principle, for instance, the classical methods for finding the complementary functions are just as good as the root-locus method of Evans.

Our preceding discussions in this chapter definitely showed that for a linear system with time-varying coefficients, the fact that all the complementary functions of the equation are damped is no guarantee for the satisfactory performance of the system with forcing functions. With some input, or forcing function, the output may even be unbounded, in spite of the damped complementary functions. Then the question of satisfactory performance of the system cannot be answered without

knowing the input function. This requirement of specifying the input, together with the difficulty of actually determining the solution of a nonhomogeneous differential equation with variable coefficients, seems to make the task of developing a general theory of stability and control for such systems a hopeless one. However, we must distinguish the computational difficulties from the real difficulty in organizing a general theory. Computational difficulties can be removed by fast electromechanical computers and should not be considered real difficulties. When this is realized, we see that specifying the input function for performance analysis is in fact designing for a specified purpose: we must know first what we want from the system under what circumstances, before we can design the system. When this approach is adopted, the general problem of stability is suppressed, because, if the system is designed to have a specified satisfactory performance, that in itself is sufficient. For linear systems with time-varying coefficients, a general theory of control design can be formulated on this basis. This is an application of the classical ballistic perturbation theory and will be discussed in the following chapter. In retrospect, we may say that the theory of conventional servomechanisms is a theory for *general design* of a *specific type* of systems. The perturbation theory of the following chapter is a theory of *specific design* of a more *general type* of systems.

CHAPTER 13

CONTROL DESIGN BY PERTURBATION THEORY

The object of the *ballistic perturbation theory* is to calculate the behavior of a projectile near the so-called normal trajectory. The normal trajectory is a certain trajectory with specified initial conditions, propulsion program, atmospheric conditions, and programed lift and drag. If the actual conditions are slightly different from these specified conditions or if the vehicle is disturbed from its normal trajectory by accidental wind gusts, the trajectory of the projectile will be different from the normal trajectory. But if such disturbance influences are small, the perturbed trajectory will lie in the neighborhood of the normal trajectory, and the difference of the perturbed trajectory and the normal trajectory will remain small. This fact of nearness to a calculated, known trajectory is the basis for the linearization of the differential equations of motion for the perturbed trajectory. The perturbed system is then represented as a linear system with time-varying coefficients, time varying because of the varying conditions of the projectile with respect to time.

The original purpose of ballistic perturbation theory was to calculate the small modification of the trajectory of a projectile due to deviations of the weight of the shell from standard value, to changes of atmospheric conditions, to effects of wind, etc. However, with the advent of modern large and fast computing machines, the tendency was to calculate all neighboring trajectories separately. The usefulness of ballistic perturbation theory then vanishes. However, the problem of designing the control for linear systems with variable coefficients is just the problem to which the ballistic perturbation theory can be applied. We shall show this in this chapter by studying the control problem of a long-range rocket vehicle. This particular problem was studied by Drenick.[1] Our treatment is, however, more complete and includes the problem of the automatic guidance of such vehicles.[2]

13.1 Equations of Motion of a Rocket. In order not to complicate matters, the vehicle is assumed to move in the equatorial plane of the

[1] R. Drenick, *J. Franklin Inst.*, **25**, 423–436 (1951).

[2] *Cf.* H. S. Tsien, T. C. Adamson, E. L. Knuth, *J. Am. Rocket Soc.*, **22**, 192–199 (1952).

rotating earth, as sketched in Fig. 13.1. The planar motion is possible due to the absence of the cross Coriolis force in the equatorial plane. The coordinate system is fixed with respect to the rotating earth, *i.e.*, it actually rotates with the angular velocity Ω, the speed of earth rotation. In the equatorial plane, the position of the vehicle at any time instant t is specified by the radius r and the angle θ from the starting point of the

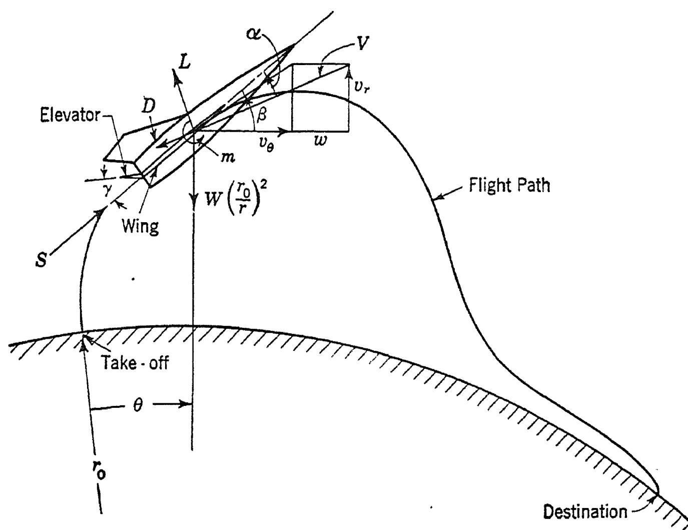


FIG. 13.1

vehicle. r_0 is the radius of the earth. g is the gravitational constant at the surface of the earth without the centrifugal force due to rotation. Let R and Θ be the force per unit mass due to thrust and aerodynamic forces acting on the vehicle in the radial and the circumferential directions, respectively. Then the equations of motion of the center of gravity of the vehicle are

$$\left.\begin{aligned} \frac{dr}{dt} &= \dot{r} \\ \frac{d\theta}{dt} &= \dot{\theta} \\ \frac{d\dot{r}}{dt} &= R + r(\dot{\theta} \pm \Omega)^2 - g\left(\frac{r_0}{r}\right)^2 \\ r\frac{d\dot{\theta}}{dt} &= \Theta - 2\dot{r}(\dot{\theta} \pm \Omega) \end{aligned}\right\} \tag{13.1}$$

where the plus sign in the second terms on the right will be valid for flights toward the east, and the minus sign for flights toward the west.

The forces R and Θ are composed of the thrust S, the lift L, and the drag D. Let W be the instantaneous weight of the vehicle with respect to g, the gravitational constant, and V the magnitude of air velocity relative to the vehicle. Then it is convenient to introduce the parameters Σ, Λ, and Δ, defined as

$$\Sigma = \frac{Sg}{W} \qquad \Lambda = \frac{Lg}{WV} \qquad \Delta = \frac{Dg}{WV} \tag{13.2}$$

It will be assumed that the natural wind velocity w is in the horizontal direction and in the equatorial plane, with positive sign if it is a head wind. w is considered to be a function of altitude r. If v_r is the radial velocity and v_θ the circumferential velocity, *i.e.*,

$$\left.\begin{aligned} v_r &= \dot{r} \\ v_\theta &= r\dot{\theta} \end{aligned}\right\} \tag{13.3}$$

then the relative air velocity V is computed as

$$V^2 = \dot{r}^2 + (r\dot{\theta} + w)^2 \tag{13.4}$$

If β is the angle between the thrust line and the horizontal direction, then the radial and the circumferential components of the forces R and Θ per unit mass are

$$\left.\begin{aligned} R &= \Sigma \sin\beta + (v_\theta + w)\Lambda - v_r\Delta \\ \Theta &= \Sigma \cos\beta - v_r\Lambda - (v_\theta + w)\Delta \end{aligned}\right\} \tag{13.5}$$

If N is the moment of forces about the center of gravity, divided by the moment of inertia of the vehicle, the equation for the angular acceleration is

$$\frac{d\dot{\beta}}{dt} = \frac{d\dot{\theta}}{dt} + N \tag{13.6}$$

To specify completely the motion of the vehicle, the lift L, the drag D, and the moment m about the center of gravity have to be given as functions of time. According to aerodynamic convention, L and D will be expressed in terms of the lift coefficient C_L and the drag coefficient C_D as follows:

$$\left.\begin{aligned} L &= \tfrac{1}{2}\rho V^2 A C_L \\ D &= \tfrac{1}{2}\rho V^2 A C_D \end{aligned}\right\} \tag{13.7}$$

where ρ is the air density, a function of the altitude r, and A is a fixed reference area, say the wing area of the vehicle. In the present problem,

since the motion of the vehicle is restricted to the equatorial plane, the attitude of the vehicle, essential for aerodynamic calculations, is determined by the angle of attack α, that is, the angle between the thrust line and the relative air velocity vector (Fig. 13.1). The control on the motion of the vehicle is effected, however, through the elevator angle γ. The parameters which will affect C_L and C_D are thus α and γ. In addition, the aerodynamic coefficients are functions of the Reynolds number Re and the Mach number M. If a is the velocity of sound in the atmosphere, a function of the altitude r, the Mach number is

$$M = \frac{V}{a} \tag{13.8}$$

If l is a typical length of the vehicle and μ is the viscosity of air, again a function of the altitude r, the Reynolds number is

$$\mathrm{Re} = \frac{\rho V l}{\mu} \tag{13.9}$$

Therefore

$$\begin{aligned} C_L &= C_L(\alpha,\gamma,M,\mathrm{Re}) \\ C_D &= C_D(\alpha,\gamma,M,\mathrm{Re}) \end{aligned} \tag{13.10}$$

We shall assume that the thrust line passes through the center of gravity of the vehicle; thus the thrust gives no moment. Since the angular motion of the vehicle during the powered flight is expected to be slow, the jet damping moment of the rocket is negligible. The only moment acting on the vehicle is then the aerodynamic moment m. m can also be expressed as a coefficient C_M as follows:

$$m = \tfrac{1}{2}\rho V^2 A l C_M \tag{13.11}$$

The moment coefficient C_M is again a function of the four variables α, γ, M, and Re, or

$$C_M = C_M(\alpha,\gamma,M,\mathrm{Re}) \tag{13.12}$$

If I is the instantaneous lateral moment of inertia of the vehicle about the center of gravity of the vehicle, then the magnitude of N in Eq. (13.6) is

$$N = \frac{m}{I} \tag{13.13}$$

With the notations defined above, the system of equations of motion is as follows:

$$
\left.
\begin{aligned}
\frac{dr}{dt} &= v_r \\
\frac{d\theta}{dt} &= \frac{v_\theta}{r} \\
\frac{d\beta}{dt} &= \dot{\beta} \\
\frac{dv_r}{dt} &= \Sigma \sin \beta + (v_\theta + w)\Lambda - v_r\Delta + r\left(\frac{v_\theta}{r} \pm \Omega\right)^2 - g\left(\frac{r_0}{r}\right)^2 = F \\
\frac{dv_\theta}{dt} &= \Sigma \cos \beta - v_r\Lambda - (v_\theta + w)\Delta - 2v_r\left(\frac{v_\theta}{r} \pm \Omega\right) + \frac{v_\theta v_r}{r} = G \\
\frac{d\dot{\beta}}{dt} &= \frac{1}{r}\left\{\Sigma \cos \beta - v_r\Lambda - (v_\theta + w)\Delta - 2v_r\left(\frac{v_\theta}{r} \pm \Omega\right)\right\} + N = H
\end{aligned}
\right\} \quad (13.14)
$$

This system of equations is a set of first-order equations for the six unknowns r, θ, β, v_r, v_θ, and $\dot{\beta}$. To solve it, the six initial values at the start, when $t = 0$, for the unknowns must be specified. In addition, the thrust S, the weight W, and the moment of inertia I must be given for every time instant. To determine the aerodynamic forces the elevator angle γ must be specified as a function of time. The properties of the atmosphere must be known; *i.e.*, the wind velocity w, air density ρ, air viscosity μ, and sound velocity a must be given as functions of the altitude r. The angle of attack α of the vehicle cannot be specified; it is a quantity to be computed from the angle β and the relative air velocity vector V.

Let the properties of the atmosphere be standardized, and the average characteristics of the vehicle and its power plant be taken to be representative. Then if the elevator angle γ is specified as a function of time, the flight path of the vehicle is determined and can be calculated by integrating the system of Eq. (13.14). The actual execution of this computation will probably be done on an electromechanical computer. This flight path of a standardized vehicle in standard atmosphere can be called the *normal flight path.*

The dominating feature of the normal flight path is its range. This range is the distance between the take-off point and the landing point. The problem of navigation is then to calculate the proper time for cutoff of the rocket motor and the proper variation of the elevator angle during flight so that the resultant range is that desired. This problem of navigation for the standardized vehicle in the standard atmosphere can be solved mathematically before the actual take-off of the vehicle, since all information for the normal flight path is known or specified beforehand.

13.2 Perturbation Equations. Natural atmospheric characteristics do not necessarily coincide with those specified for the standard atmosphere. The wind velocity at each altitude changes according to the weather conditions; the temperature T is also a varying quantity. Therefore one should expect variations from the normal flight path due to changes in atmospheric conditions. The actual vehicle also may be somewhat different from the standardized vehicle in weight, in rocket motor performance, etc. Therefore the actual flight path will be different from the normal flight path if the elevator angle program of the normal path is used. The problem of navigation of an actual vehicle is that of correcting the elevator angle program so that the range of the actual flight will be the same as the normal flight path and the destination is reached without error. Because of the rapidity of flight, this navigational problem cannot be solved by the conventional method, which neglects completely the dynamic effects and is based upon only kinematical considerations. But instead, the problem should be solved by an automatic computing system, which responds to every deviation from the normal conditions with a speed approaching instant action. The problem is thus more appropriately called the guidance problem, and the control system, the guidance system.

The general problem of guidance is very difficult indeed. However, the deviations from the normal conditions are expected to be small, since the normal flight path is, after all, a good representation of the average situation. This fact immediately suggests that only first-order quantities in deviations need be considered. This "linearization" is the basis of the ballistic perturbation theory. The resultant system will have coefficients that are evaluated on the normal flight path and are generally functions of time. Therefore the fundamental equations in the ballistic perturbation theory are linear equations with time-varying coefficients. Our discussion of the guidance problem of a long-range rocket vehicle is thus an example of control design for such systems. The particular design specification here is the vanishing of range error. The controlled "input" here is the elevator angle corrections. We shall develop these concepts in the following discussion.

Let the quantities of the normal flight path be denoted by a bar over them, and deviations, by the δ sign. Thus for the actual flight path,

$$\left.\begin{array}{lll} r = \bar{r} + \delta r & \theta = \bar{\theta} + \delta\theta & \beta = \bar{\beta} + \delta\beta \\ v_r = \bar{v}_r + \delta v_r & v_\theta = \bar{v}_\theta + \delta v_\theta & \dot{\beta} = \dot{\bar{\beta}} + \delta\dot{\beta} \end{array}\right\} \qquad (13.15)$$

The deviations of the actual atmosphere from the standard atmosphere are expressed as the deviation of density $\delta\rho$, the deviation of temperature δT, and the deviation of wind velocity δw; thus

$$\rho = \bar{\rho} + \delta\rho \qquad T = \bar{T} + \delta T \qquad w = \bar{w} + \delta w \qquad (13.16)$$

If we assume no change in the composition of the atmosphere at any altitude from that of the standard atmosphere, the knowledge of $\delta\rho$ and δT is sufficient to calculate the deviation of pressure, if necessary. The deviation of the actual vehicle from the normal vehicle is assumed to be limited only to the deviation of weight δW and the deviation of moment of inertia δI. That is,

$$W = \bar{W} + \delta W \qquad I = \bar{I} + \delta I \tag{13.17}$$

The thrust S is assumed to be fixed at the standard value. The wing area A and the aerodynamic characteristics of the vehicle, as expressed by Eqs. (13.10) and (13.12), are also assumed to be invariant.

Substituting Eqs. (13.15) to (13.17) into Eq. (13.14) and then subtracting the corresponding equation for the normal flight path, we have the following equations, according to our linearizing principle:

$$\left.\begin{aligned} \frac{d\,\delta r}{dt} &= \delta v_r \\ \frac{d\,\delta\theta}{dt} &= -\frac{\bar{v}_\theta}{\bar{r}^2}\,\delta r + \frac{1}{\bar{r}}\,\delta v_\theta \\ \frac{d\,\delta\beta}{dt} &= \delta\dot{\beta} \end{aligned}\right\} \tag{13.18}$$

$$\left.\begin{aligned} \frac{d\,\delta v_r}{dt} &= a_1\,\delta r + a_2\,\delta\beta + a_3\,\delta v_r + a_4\,\delta v_\theta + a_5\,\delta\gamma + a_6\,\delta\rho \\ &\qquad + a_7\,\delta T + a_8\,\delta w + a_9\,\delta W \\ \frac{d\,\delta v_\theta}{dt} &= b_1\,\delta r + b_2\,\delta\beta + b_3\,\delta v_r + b_4\,\delta v_\theta + b_5\,\delta\gamma + b_6\,\delta\rho \\ &\qquad + b_7\,\delta T + b_8\,\delta w + b_9\,\delta W \\ \frac{d\,\delta\dot{\beta}}{dt} &= c_1\,\delta r + c_2\,\delta\beta + c_3\,\delta v_r + c_4\,\delta v_\theta + c_5\,\delta\gamma + c_6\,\delta\rho \\ &\qquad + c_7\,\delta T + c_8\,\delta w + c_9\,\delta W + c_{10}\,\delta I \end{aligned}\right\} \tag{13.19}$$

The coefficient a's, b's, and c's are partial derivatives of the F, G, and H defined by Eq. (13.14), evaluated on the normal flight path. For example,

$$\left.\begin{aligned} a_1 &= \overline{\left(\frac{\partial F}{\partial r}\right)} & a_2 &= \overline{\left(\frac{\partial F}{\partial \beta}\right)} & a_3 &= \overline{\left(\frac{\partial F}{\partial v_r}\right)} \\ a_4 &= \overline{\left(\frac{\partial F}{\partial v_\theta}\right)} & a_5 &= \overline{\left(\frac{\partial F}{\partial \gamma}\right)} & a_6 &= \overline{\left(\frac{\partial F}{\partial \rho}\right)} \\ a_7 &= \overline{\left(\frac{\partial F}{\partial T}\right)} & a_8 &= \overline{\left(\frac{\partial F}{\partial w}\right)} & a_9 &= \overline{\left(\frac{\partial F}{\partial W}\right)} \end{aligned}\right\} \tag{13.20}$$

The details of this calculation of coefficients are given in the appendix to this chapter.

Equations (13.18) and (13.19) are linear equations with variable coefficients for the six deviation quantities. If the deviations of the atmospheric properties $\delta\rho$, δT, and δw are known, and if $\delta\gamma$, δW, and δI are specified, then this system of differential equations determines δr, $\delta\theta$, $\delta\beta$, δv_r, δv_θ, and $\delta\dot{\beta}$. The problem of guidance is, however, different from this. What is required is the function $\delta\gamma$, or the correction to the elevator angle, such that the range error is zero. As suggested by Drenick, this guidance problem can best be solved by the method of *adjoint functions* of Bliss.[1]

13.3 Adjoint Functions. The principle of the method of adjoint functions is as follows. Let $y_i(t)$, where $i = 1, \ldots, n$, be determined by a system of n linear equations

$$\frac{dy_i}{dt} - \sum_{j=1}^{n} a_{ij}y_j = Y_i(t) \qquad \text{for } i = 1, \ldots, n \tag{13.21}$$

where a_{ij} are given coefficients which may be functions of the time t. The $Y_i(t)$ are the "forcing" functions or inputs. Now introduce a new set of functions $\lambda_i(t)$, where $i = 1, \ldots, n$, called the adjoint functions to $y_i(t)$, which satisfy the following system of homogeneous equations

$$\frac{d\lambda_i}{dt} + \sum_{j=1}^{n} a_{ji}\lambda_j = 0 \qquad i = 1, \ldots, n \tag{13.22}$$

By multiplying Eq. (13.21) by λ_i and Eq. (13.22) by y_i and summing the equations over i, we have

$$\frac{d}{dt}\sum_{i=1}^{n} \lambda_i y_i - \sum_{i=1}^{n}\sum_{j=1}^{n} (a_{ij}\lambda_i y_j - a_{ji}\lambda_j y_i) = \sum_{i=1}^{n} \lambda_i Y_i$$

The two parts in the double sum evidently cancel each other, and we have

$$\frac{d}{dt}\sum_{i=1}^{n} \lambda_i y_i = \sum_{i=1}^{n} \lambda_i Y_i \tag{13.23}$$

This equation can be integrated from the instant $t = t_1$ to the instant $t = t_2$, and

$$\sum_{i=1}^{n} \lambda_i y_i \Big|_{t=t_2} = \sum_{i=1}^{n} \lambda_i y_i \Big|_{t=t_1} + \int_{t_1}^{t_2} \left(\sum_{i=1}^{n} \lambda_i y_i\right) dt \tag{13.24}$$

Bliss calls this equation the *fundamental formula.*

[1] G. A. Bliss, "Mathematics for Exterior Ballistics," John Wiley & Sons, Inc., New York, 1944.

For the problem of long-range rocket, the y_i are the perturbation quantities, that is, $n = 6$ and

$$\left.\begin{array}{lll} y_1 = \delta r & y_2 = \delta\theta & y_3 = \delta\beta \\ y_4 = \delta v_r & y_5 = \delta v_\theta & y_6 = \delta\dot{\beta} \end{array}\right\} \tag{13.25}$$

Then, according to Eq. (13.19), the adjoint functions $\lambda_i(t)$ satisfy the following system of equations

$$\left.\begin{array}{llll} -\dfrac{d\lambda_1}{dt} = & -\dfrac{\bar{v}_\theta}{\bar{r}^2}\lambda_2 & + a_1\lambda_4 + b_1\lambda_5 + c_1\lambda_6 \\ -\dfrac{d\lambda_2}{dt} = 0 & & \\ -\dfrac{d\lambda_3}{dt} = & & a_2\lambda_4 + b_2\lambda_5 + c_2\lambda_6 \\ -\dfrac{d\lambda_4}{dt} = \lambda_1 & & + a_3\lambda_4 + b_3\lambda_5 + c_3\lambda_6 \\ -\dfrac{d\lambda_5}{dt} = & \dfrac{1}{\bar{r}}\lambda_2 & + a_4\lambda_4 + b_4\lambda_5 + c_4\lambda_6 \\ -\dfrac{d\lambda_6}{dt} = & \lambda_3 & \end{array}\right\} \tag{13.26}$$

The inputs Y_i are

$$Y_1 = Y_2 = Y_3 = 0 \tag{13.27}$$

$$\left.\begin{array}{l} Y_4 = a_5\,\delta\gamma + a_6\,\delta p + a_7\,\delta T + a_8\,\delta w + a_9\,\delta W \\ Y_5 = b_5\,\delta\gamma + b_6\,\delta p + b_7\,\delta T + b_8\,\delta w + b_9\,\delta W \\ Y_6 = c_5\,\delta\gamma + c_6\,\delta p + c_7\,\delta T + c_8\,\delta w + c_9\,\delta W + c_{10}\,\delta I \end{array}\right\} \tag{13.28}$$

13.4 Range Correction. Equation (13.26) does not determine the λ functions completely. To do that, a set of values for λ must be specified at a certain instant. What values to pick for λ at what instant depends upon the specific purpose of the control design. For our guidance problem, we require zero range error. Therefore the quantity of interest is $\delta\theta$ at the instant of landing, or $\delta\theta_2$ if the subscript $_2$ denotes quantities at the instant of landing. This is sufficient to determine the λ's completely. We shall see this presently.

If t_2 is the time instant of landing of the actual vehicle and $\bar{t}_2$ the time instant of landing of the normal flight path, then

$$t_2 = \bar{t}_2 + \delta t_2 \tag{13.29}$$

Similarly

$$\left.\begin{array}{l} r_2 = \bar{r}_2 + \delta r_2 \\ \theta_2 = \bar{\theta}_2 + \delta\theta_2 \end{array}\right\} \tag{13.30}$$

but

$$\left.\begin{aligned} \delta r_2 &= (\bar{v}_r)_{t=\bar{t}_2}\,\delta t_2 + (\delta r)_{t=\bar{t}_2} \\ \delta\theta_2 &= \frac{1}{r_0}\,(\bar{v}_\theta)_{t=\bar{t}_2}\,\delta t_2 + (\delta\theta)_{t=\bar{t}_2} \end{aligned}\right\} \tag{13.31}$$

However, δr_2 is by definition zero, because landing means contact with the surface of earth, or $r_2 = \bar{r}_2 = r_0$. By eliminating δt_2 from Eq. (13.31),

$$\delta\theta_2 = \left[-\frac{1}{\bar{r}}\left(\frac{\bar{v}_\theta}{\bar{v}_r}\right)\delta r + \delta\theta\right]_{t=\bar{t}_2} \tag{13.32}$$

Therefore, if the magnitudes of λ_i at the landing instant $t = \bar{t}_2$ are specified as

$$\left.\begin{aligned} \lambda_1 &= -\frac{1}{\bar{r}}\left(\frac{\bar{v}_\theta}{\bar{v}_r}\right) \qquad \lambda_2 = 1 \\ \lambda_3 &= \lambda_4 = \lambda_5 = \lambda_6 = 0 \end{aligned}\right\} \tag{13.33}$$

then the error in range is given by

$$\begin{aligned} \delta\theta_2 &= \sum_{i=1}^{n} \lambda_i y_i \Bigg|_{t=\bar{t}_2} \\ &= [\lambda_1\,\delta r + \lambda_2\,\delta\theta + \lambda_3\,\delta\beta + \lambda_4\,\delta v_r + \lambda_5\,\delta v_\theta + \lambda_6\,\delta\dot{\beta}]_{t=\bar{t}_2} \end{aligned} \tag{13.34}$$

When the normal flight path is determined, the coefficients in Eq. (13.27) are specified as functions of time. These equations together

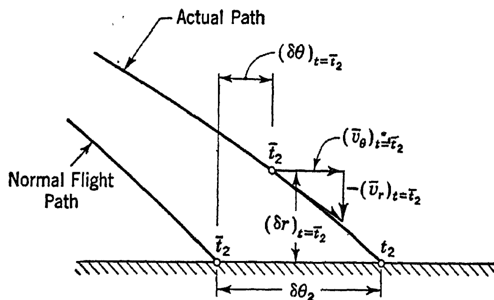


Fig. 13.2

with the end conditions of Eq. (13.33) then determine the adjoint functions λ_i. The integration has to be performed "backwards" for $t < \bar{t}_2$, perhaps by an electromechanical computer. With the adjoint functions so determined, one can use the fundamental formula of Eq. (13.25) to modify the equation for the range error given by Eq. (13.34): Let $\bar{t}_1$ denote the time instant for the power cutoff for the normal flight path. Then the condition for the error in range $\delta\theta_2$ to be zero can be expressed as

$$\delta\theta_2 = 0 = [\lambda_1\,\delta r + \lambda_2\,\delta\theta + \lambda_3\,\delta\beta + \lambda_4\,\delta v_r + \lambda_5\,\delta v_\theta + \lambda_6\,\delta\dot{\beta}]_{t=\bar{t}_1} + \int_{t=\bar{t}_1}^{t=\bar{t}_2} [\lambda_4 Y_4 + \lambda_5 Y_5 + \lambda_6 Y_6]\,dt \quad (13.35)$$

This is the basic equation for guidance. It will be exploited in the following sections.

13.5 Cutoff Condition. The condition of Eq. (13.35) for arbitrary disturbances can be broken down into two parts; the sum and the integral are set equal to zero separately. Therefore the condition to be satisfied at the normal cutoff instant $\bar{t}_1$ is

$$[\lambda_1\,\delta r + \lambda_2\,\delta\theta + \lambda_3\,\delta\beta + \lambda_4\,\delta v_r + \lambda_5\,\delta v_\theta + \lambda_6\,\delta\dot{\beta}]_{t=\bar{t}_1} = 0 \quad (13.36)$$

Since the normal cutoff instant $\bar{t}_1$ is a standard time instant, but not necessarily the actual cutoff instant t_1, *i.e.*,

$$t_1 = \bar{t}_1 + \delta t_1 \quad (13.37)$$

Eq. (13.36) should be converted into a more useful form involving the quantities at the actual cutoff instant. This is easily done, because up to the first-order quantities, according to Eq. (13.35),

$$(\delta r)_{t=\bar{t}_1} = (r)_{t=t_1} - \overline{\left(\frac{dr}{dt}\right)}_{t=\bar{t}_1} \delta t_1 - (\bar{r})_{t=\bar{t}_1}$$

or

$$(\delta r)_{t=\bar{t}_1} = (r)_{t=t_1} - (\bar{r})_{t=\bar{t}_1} - (\bar{v}_r)_{t=\bar{t}_1}\,\delta t_1$$

Similarly,

$$(\delta\theta)_{t=\bar{t}_1} = (\theta)_{t=t_1} - (\bar{\theta})_{t=\bar{t}_1} - \left(\frac{1}{\bar{r}}\,\bar{v}_\theta\right)_{t=\bar{t}_1} \delta t_1$$

$$(\delta\beta)_{t=\bar{t}_1} = (\beta)_{t=t_1} - (\bar{\beta})_{t=\bar{t}_1} - (\bar{\dot{\beta}})_{t=\bar{t}_1}\,\delta t_1$$

$$(\delta v_r)_{t=\bar{t}_1} = (v_r)_{t=t_1} - (\bar{v}_r)_{t=\bar{t}_1} - (\bar{F})_{t=\bar{t}_1}\,\delta t_1$$

$$(\delta v_\theta)_{t=\bar{t}_1} = (v_\theta)_{t=t_1} - (\bar{v}_\theta)_{t=\bar{t}_1} - (\bar{G})_{t=\bar{t}_1}\,\delta t_1$$

$$(\delta\dot{\beta})_{t=\bar{t}_1} = (\dot{\beta})_{t=t_1} - (\bar{\dot{\beta}})_{t=\bar{t}_1} - (\bar{H})_{t=\bar{t}_1}\,\delta t_1$$

where $\bar{F}$, $\bar{G}$, and $\bar{H}$ are the values of these quantities, given by Eq. (13.14), evaluated on the normal flight path. In fact they should be evaluated at an instant just before the normal cutoff time $\bar{t}_1$ so that the accelerating force of the rocket is included and the rates of change of velocities are those of a powered flight. Now define J and $\bar{J}$ as follows:

$$J = [\lambda_1^* r + \lambda_2^* \theta + \lambda_3^* \beta + \lambda_4^* v_r + \lambda_5^* v_\theta + \lambda_6^* \dot{\beta}]_{t=t_1} \quad (13.38)$$

and

$$\bar{J} = [\lambda_1^* \bar{r} + \lambda_2^* \bar{\theta} + \lambda_3^* \bar{\beta} + \lambda_4^* \bar{v}_r + \lambda_5^* \bar{v}_\theta + \lambda_6^* \bar{\dot{\beta}}]_{t=\bar{t}_1}$$

where λ_i^* are the values of λ_i evaluated at the normal cutoff time $\bar{t}_1$. Then the condition to be satisfied at the actual cutoff instant t_1 is

$$J = \bar{J} + \left[\lambda_1^* \bar{v}_r + \lambda_2^* \frac{\bar{v}_\theta}{\bar{r}} + \lambda_3^* \bar{\beta} + \lambda_4^* \bar{F} + \lambda_5^* \bar{G} + \lambda_6^* \bar{H} \right]_{t=\bar{t}_1} (t_1 - \bar{t}_1) \quad (13.39)$$

This is the equation for determining the proper instant of power cutoff.

When the normal flight path is known, $\bar{J}$ and the quantity within the bracket to the right in Eq. (13.39) are fixed. Then the whole right-hand side of Eq. (13.39) can be considered as a linearly increasing function of time t if t is substituted for t_1. Simultaneously J can be computed at every instant before cutoff by using the predetermined λ_i^* and the values of position and velocity of the actual vehicle obtained by tracking stations. The magnitudes of the quantities on the two sides of Eq. (13.39) can then be continuously compared. When they are equal to each other, Eq. (13.39) is satisfied. Then the power cutoff signal is given, and the rocket power is shut off.

13.6 Guidance Condition. When the rocket power is shut off earlier or later than the normal cutoff instant $\bar{t}_1$, the propellant left in the tank, if not dumped, will alter the weight W and the moment of inertia I of the vehicle. It is also possible that the pay-load of the vehicle is not that specified for the standard vehicle. Then, after power cutoff, there is a fixed δW and δI, fixed in the sense that they do not change with time and are known once the power cutoff is effected. Of a different character are the deviations $\delta\rho$, δT, and δw of the actual atmosphere from the standard atmosphere. These are not known unless they are measured. In the following, it is proposed to use the vehicle itself as a measuring instrument, and we proceed as follows.

After the cutoff condition is satisfied, the condition for zero range error is that the integral in Eq. (13.35) should vanish. Now since the Y_i in that integrand involve arbitrary disturbances $\delta\rho$, δT, and δw not known beforehand, this condition can be satisfied only if the integrand itself vanishes. That is, according to Eqs. (13.28),

$$\begin{aligned}(\lambda_4 a_5 + \lambda_5 b_5 + \lambda_6 c_5)\,\delta\gamma &+ (\lambda_4 a_6 + \lambda_5 b_6 + \lambda_6 c_6)\,\delta\rho \\ &+ (\lambda_4 a_7 + \lambda_5 b_7 + \lambda_6 c_7)\,\delta T + (\lambda_4 a_8 + \lambda_5 b_8 + \lambda_6 c_8)\,\delta w \\ &+ (\lambda_4 a_9 + \lambda_5 b_9 + \lambda_6 c_9)\,\delta W + \lambda_6 c_{10}\,\delta I = 0\end{aligned}$$

or, with the following notation

$$\left.\begin{aligned} d_5 &= \lambda_4 a_5 + \lambda_5 b_5 + \lambda_6 c_5 \\ d_6 &= \lambda_4 a_6 + \lambda_5 b_6 + \lambda_6 c_6 \\ d_7 &= \lambda_4 a_7 + \lambda_5 b_7 + \lambda_6 c_7 \\ d_8 &= \lambda_4 a_8 + \lambda_5 b_8 + \lambda_6 c_8 \\ D &= -(\lambda_4 a_9 + \lambda_5 b_9 + \lambda_6 c_9)\,\delta W - \lambda_6 c_{10}\,\delta I \end{aligned}\right\} \quad (13.40)$$

this condition can be written as

$$d_5\,\delta\gamma + d_6\,\delta\rho + d_7\,\delta T + d_8\,\delta w = D \quad (13.41)$$

Equation (13.19) can be rewritten as

$$\left.\begin{aligned} a_5\,\delta\gamma + a_6\,\delta\rho + a_7\,\delta T + a_8\,\delta w &= A \\ b_5\,\delta\gamma + b_6\,\delta\rho + b_7\,\delta T + b_8\,\delta w &= B \\ c_5\,\delta\gamma + c_6\,\delta\rho + c_7\,\delta T + c_8\,\delta w &= C \end{aligned}\right\} \tag{13.42}$$

where

$$\left.\begin{aligned} A &= \frac{d\,\delta v_r}{dt} - a_1\,\delta r - a_2\,\delta\beta - a_3\,\delta v_r - a_4\,\delta v_\theta - a_9\,\delta W \\ B &= \frac{d\,\delta v_\theta}{dt} - b_1\,\delta r - b_2\,\delta\beta - b_3\,\delta v_r - b_4\,\delta v_\theta - b_9\,\delta W \\ C &= \frac{d\,\delta\dot{\beta}}{dt} - c_1\,\delta r - c_2\,\delta\beta - c_3\,\delta v_r - c_4\,\delta v_\theta - c_9\,\delta W - c_{10}\,\delta I \end{aligned}\right\} \tag{13.43}$$

If the tracking stations for the vehicle will measure the quantities A, B, and C, then the atmospheric disturbances $\delta\rho$, δT, and δw can be determined by solving for these variations using Eq. (13.42). This is essentially using the vehicle itself as a measuring instrument for $\delta\rho$, δT, and δw. When $\delta\rho$, δT, and δw are known, Eq. (13.41) gives the proper elevator angle correction $\delta\gamma$.

A mathematically equivalent way to calculate $\delta\gamma$ would be to solve directly for $\delta\gamma$ using the system of Eqs. (13.41) and (13.42). Thus

$$\begin{vmatrix} a_5 & a_6 & a_7 & a_8 \\ b_5 & b_6 & b_7 & b_8 \\ c_5 & c_6 & c_7 & c_8 \\ d_5 & d_6 & d_7 & d_8 \end{vmatrix} \delta\gamma = \begin{vmatrix} A & a_6 & a_7 & a_8 \\ B & b_6 & b_7 & b_8 \\ C & c_6 & c_7 & c_8 \\ D & d_6 & d_7 & d_8 \end{vmatrix} \tag{13.44}$$

This equation specifies the necessary change in the elevator angle at every instant, to be calculated from a's, b's, c's, and A, B, C, D at the same instant. These quantities consist partly of predetermined information from the normal flight path, and partly of measured information on the position and the velocities of vehicle obtained by tracking the vehicle. At high altitudes where the air density is very small, the aerodynamic forces will be almost negligible in comparison with the gravitational and inertial forces. Then the quantities A, B, and C of Eq. (13.43) will be the small difference of large magnitudes. These are then the quantities most difficult to determine accurately. If the actual elevator angle is made to conform with the one calculated by Eq. (13.44), then in conjunction with the proper power cutoff, as specified in the last section, the vehicle will be navigated to the chosen landing point in spite of the atmospheric disturbances.

13.7 Guidance System. When the general character of the flight path has been chosen from over-all engineering considerations, the first step is the calculation of the normal flight path using the properties of

the standard atmosphere and the expected performance of the vehicle with normal weight. The knowledge of the normal flight path then determines the a's, b's, and c's. Equation (13.27) together with the end conditions of Eq. (13.33) allows the calculation of the adjoint functions λ_i. All this information should be on hand before the actual flight of the vehicle and may be called the "stored data."

Before the power cutoff, the elevator angle may be programed according to that for the normal flight path, and the stability of the vehicle is supplied by the jet vanes or by the auxiliary rockets. The tracking stations, however, go immediately into action and supply the vehicle with information on its positions and velocities. This information goes first into the cutoff computer which, using the stored information, continuously compares the magnitudes of quantities on the two sides of Eq. (13.39), the cutoff condition. When that condition is satisfied, the power cutoff is effected.

At the instant of power cutoff, the tracking information is switched to the computer for the guidance system. The instant of power cutoff also fixes the amount of propellant in the tank and thus determines the variations of the weight W and the moment of inertia I from the normal. This information together with the stored data on the normal flight path then allows the computer to generate the elevator correction angle $\delta\gamma$ according to Eqs. (13.40), (13.43), and (13.44). Theoretically the value of $\delta\gamma$ must be obtained without time delay from the instant when the information is received, because Eq. (13.44) is a condition of equality of two quantities evaluated at identical time instants. The computed correction $\delta\gamma$ combined with the elevator angle $\bar{\gamma}$, determined for the normal flight path, then gives the actual elevator angle setting γ. The design of the control mechanism for the elevator from here on can follow the practice of the conventional feedback servomechanism, with the usual criteria of quick action, stability, and accuracy. The general scheme of the guidance system can then be represented by Fig. 13.3.

The computers envisaged here are carried in the vehicle and receive the information on positions and velocities of the vehicle from the fixed ground tracking stations along the flight path. Then, as indicated in Fig. 13.3, this is the feedback link. Properly designed computers here assure us of the specified performance of the system and thus function as the compensating circuit or the amplifier of the conventional servomechanism. In over-all conception then, our guidance system is very similar to the simple servomechanism studied in the preceding chapters. However, the guidance system is a system of great complexity. Its design requires the theory of perturbation together with the concept of adjoint functions. Our example of the guidance of a long-range rocket, although somewhat oversimplified, serves the purpose of showing how

the theory can be used to design the control system. In the example, there is only one design criterion, *i.e.*, the vanishing of range error. In more complicated systems, more than one design criterion can be specified, and more than one set of adjoint functions will be required. Nevertheless, the design principle will still be the same as that shown in the simple example.

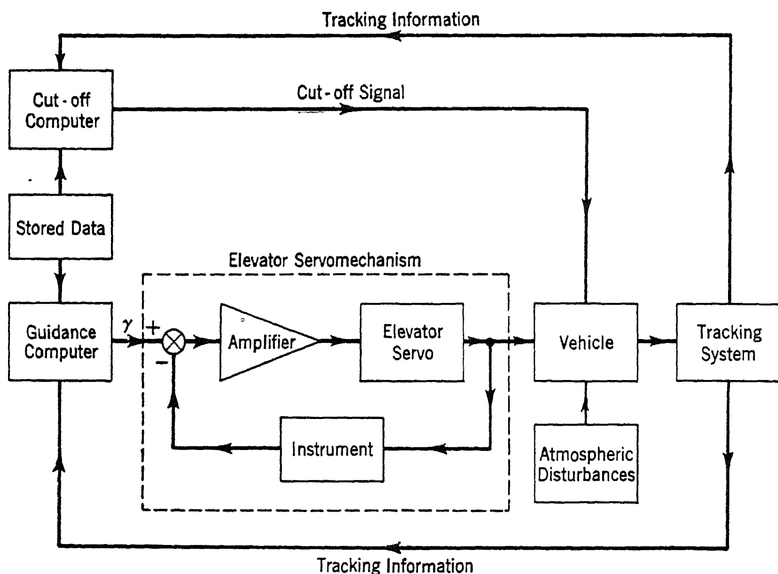


FIG. 13.3

13.8 Control Computers. Although it is not the purpose here to discuss the detailed construction and the engineering of any components of the systems considered, the role of the computer, first introduced in connection with optimum switching in Chap. 10, is so important in the more advanced control systems that a general discussion of its characteristics and requirements would be appropriate. For details, the reader should refer to books on this special subject.[1]

There are in current use two different kinds of computers: the *analog computer* and the *digital computer*. The analog computer is just what its name implies: a physical analogy to the type of problem its designer wishes to solve. It is then a system which is described by the same mathematical formalism as the computation to be carried out. The

[1] Engineering Research Associates, "High-speed Computing Devices," McGraw-Hill Book Company, Inc., 1950. For d-c analog computers, see G. A. Korn and T. M. Korn, "Electronic Analog Computers," McGraw-Hill Book Company, Inc., New York, 1952.

inputs to the machine are in terms of the value of some physical quantity —an electric voltage or current, the degree of angular rotation of a shaft, or the amount of compression of a spring. The machine transforms these inputs into other physical quantities, the outputs, according to the rules of its construction, chosen by the designer to represent the specified mathematical procedure. The inputs to an analog computer are then the instrument readings of the several physical quantities of the system to be controlled, and the outputs of the computer are command signals fed directly to the individual servomechanisms of the controlled quantities.

In contrast to the analog computer, a digital computer works by counting. Data on the problem must be supplied in the form of numbers; the machine processes this information according to the rules of arithmetic or other formal logic demanded by the computing problem, and expresses the final result in numerical form. There are two very important consequences of this manner of operation. First, input and output equipment, or the "*transducers*," must be designed to make an appropriate translation between the logical world of the digital machine and the physical world of the controlled system. Second, the problem to be solved must be formulated explicitly for the digital computer. In the case of the analog computer, the problem is implicit in the construction of the machine itself; construction of a digital computer is determined not by any particular problem or class of problems but by the logical rules which the machine must follow in the solution of the particular computing problem.

In comparing digital and analog computers as components in a control system, we observe, first, that for simple control applications the analog machine is almost always less elaborate than a digital machine would be. Even the most elementary digital computer requires an arithmetical unit, a storage unit, a control unit, and input and output transducers. For simple problems, this array of equipment is wastefully elaborate. In contrast, an analog computer need be no more complicated than the problem demands. In fact, as mentioned before, the compensating circuit in a simple servomechanism is such an analog computer.

As the computing task becomes more complex, as for the guidance problem discussed in this chapter, the analog machine loses its advantage, and we see a second fundamental difference between the two types of machine. The analog computer, being a physical analogy to the problem, must be more complicated for more complicated computing problems. If it is mechanical, longer and ever longer trains of gears, ball-and-disk integrators, and other devices must be connected together; if it is electric, more and more amplifiers must be cascaded. In the mechanical case, the inevitable looseness in the gears and linkages, though tolerable in

simple setups, will eventually add up to the point where the total backlash or "play" in the machine is bigger than the significant output quantities, and the device becomes useless. In the electric case, the random electric disturbances, the noise, which always occur in electric circuits, will similarly build up until they overwhelm the desired signals. Since noise is far less obtrusive than backlash, electric analog computers can be more complicated than their mechanical counterparts, but still there is a limit. The digital computer, on the other hand, is entirely free of the hazards of backlash and noise. There is no intrinsic limit to the complexity of the problem that a digital machine can handle.

The third important difference between analog and digital computers is in their potential accuracy. The precision of the analog computer is restricted by the accuracy with which physical quantities can be handled and measured and by the accuracy of representing a physical system by idealized laws. In practice, the best such a machine can achieve is an accuracy of about 1 part in 10,000; many give results accurate to only 1 or 2 parts in 100. For some applications this range of precision is adequate; for others it is not. On the other hand, a digital computer, which deals only with numbers, can be as precise as we wish to make it. To increase accuracy we need only increase the number of significant figures carried by the machine to represent each quantity being handled. Of course, the over-all accuracy is still limited by the accuracy of the input and output transducers; but this does not alter the fact that where high precision is required, a digital computer is preferable to the analog computer.

There is a fourth aspect in which the two types of computers differ. An analog machine works in what is called *real time*. That is, it continuously offers a solution of the problem it is solving, and this solution is appropriate at every instant to all the inputs which have so far entered the machine. On the other hand, a digital machine works by formulating and solving an explicit logical model of the computing problem. Therefore a digital machine gives only a sequence of values of the output at discrete time instants, and, furthermore, because of the finite, although short, computing time, the output lags behind the input. Two problems then arise: the problem of interpolation of the output between the discrete time instants, and the problem of prediction of the output from fast values to remove the time lag of the output. Obviously, if the computing time is very much shorter than the characteristic time of the controlled system, the prediction question can be ignored, and the computer considered to be working in real time. The present-day electronic digital computer seems to be fast enough in this respect for the guidance problems of the long-range rocket discussed previously; but for a high-speed guided missile, this time-lag effect must be properly included in the design of the control system.

Appendix to Chapter 13

CALCULATION OF PERTURBATION COEFFICIENTS

The quantities F, G, and H are defined by Eq. (13.14). They contain the parameters Σ, Λ, Δ, and N. According to their definition, as given by Eqs. (13.2) and (13.13), they can be written as follows:

$$\left.\begin{aligned} \Sigma &= \frac{Sg}{W} \\ \Lambda &= \frac{g}{W}\frac{1}{2}\rho A C_L \sqrt{v_r^2 + (v_\theta + w)^2} \\ \Delta &= \frac{g}{W}\frac{1}{2}\rho A C_D \sqrt{v_r^2 + (v_\theta + w)^2} \\ N &= \frac{1}{I}\frac{1}{2}\rho A C_M \{v_r^2 + (v_\theta + w)^2\} \end{aligned}\right\} \tag{13.45}$$

where the aerodynamic coefficients C_L, C_D, and C_M are functions of the angle of attack α, the elevator angle γ, the Mach number M, and the Reynolds number Re. These aerodynamic parameters are related to quantities immediately connected with the flight path as follows:

$$\alpha = \beta - \tan^{-1}\left(\frac{v_r}{v_\theta + w}\right) \qquad M = \frac{V}{a(r)} \qquad \mathrm{Re} = \frac{\rho V l}{\mu(r)} \tag{13.46}$$

where $a(r)$ is the sound velocity in the atmosphere, and $\mu(r)$ is the coefficient of viscosity of air, both functions of altitude r. In the following calculation, the thrust S will be considered to be a function of altitude only. It is also assumed that the composition of atmosphere at different altitudes remains the same as that of the standard atmosphere; only the density ρ and the temperature T changes. Thus the variations of a and μ at any altitude are variations due to temperature T.

For Σ:

$$\frac{\partial \Sigma}{\partial r} = \frac{g}{W}\frac{\partial S}{\partial r} \qquad \frac{\partial \Sigma}{\partial W} = -\frac{\Sigma}{W} \tag{13.47}$$

All other partial derivatives are zero.

For Λ:

$$\left.\begin{aligned} \frac{\partial \Lambda}{\partial r} &= \Lambda\left\{\frac{1}{\rho}\frac{d\rho}{dr}\left(1 + \frac{\mathrm{Re}}{C_L}\frac{\partial C_L}{\partial \mathrm{Re}}\right) + \frac{1}{V^2}\frac{dw}{dr}\left[\left(\frac{M}{C_L}\frac{\partial C_L}{\partial M} + \frac{\mathrm{Re}}{C_L}\frac{\partial C_L}{\partial \mathrm{Re}} + 1\right)\right.\right. \\ &\qquad \left.\left.\cdot (v_\theta + w) + \frac{1}{C_L}\frac{\partial C_L}{\partial \alpha} v_r\right] - \frac{M}{C_L}\frac{\partial C_L}{\partial M}\frac{1}{a}\frac{da}{dr} - \frac{\mathrm{Re}}{C_L}\frac{\partial C_L}{\partial \mathrm{Re}}\frac{1}{\mu}\frac{d\mu}{dr}\right\} \\ \frac{\partial \Lambda}{\partial v_r} &= \Lambda\frac{v_r}{V^2}\left(\frac{M}{C_L}\frac{\partial C_L}{\partial M} + \frac{\mathrm{Re}}{C_L}\frac{\partial C_L}{\partial \mathrm{Re}} + 1 - \frac{1}{C_L}\frac{\partial C_L}{\partial \alpha}\frac{v_\theta + w}{v_r}\right) \\ \frac{\partial \Lambda}{\partial v_\theta} &= \Lambda\frac{v_\theta + w}{V^2}\left(\frac{M}{C_L}\frac{\partial C_L}{\partial M} + \frac{\mathrm{Re}}{C_L}\frac{\partial C_L}{\partial \mathrm{Re}} + 1 + \frac{1}{C_L}\frac{\partial C_L}{\partial \alpha}\frac{v_r}{v_\theta + w}\right) \\ \frac{\partial \Lambda}{\partial \beta} &= \Lambda\frac{1}{C_L}\frac{\partial C_L}{\partial \alpha} \\ \frac{\partial \Lambda}{\partial \gamma} &= \Lambda\frac{1}{C_L}\frac{\partial C_L}{\partial \gamma} \\ \frac{\partial \Lambda}{\partial \rho} &= \Lambda\frac{1}{\rho}\left(1 + \frac{\mathrm{Re}}{C_L}\frac{\partial C_L}{\partial \mathrm{Re}}\right) \\ \frac{\partial \Lambda}{\partial T} &= -\Lambda\left(\frac{M}{C_L}\frac{\partial C_L}{\partial M}\frac{1}{2T} + \frac{\mathrm{Re}}{C_L}\frac{\partial C_L}{\partial \mathrm{Re}}\frac{1}{\mu}\frac{\partial \mu}{\partial T}\right) \\ \frac{\partial \Lambda}{\partial w} &= \Lambda\frac{v_\theta + w}{V^2}\left(\frac{M}{C_L}\frac{\partial C_L}{\partial M} + \frac{\mathrm{Re}}{C_L}\frac{\partial C_L}{\partial \mathrm{Re}} + 1 + \frac{1}{C_L}\frac{\partial C_L}{\partial \alpha}\frac{v_r}{v_\theta + w}\right) = \frac{\partial \Lambda}{\partial v_\theta} \\ \frac{\partial \Lambda}{\partial W} &= -\frac{\Lambda}{W} \end{aligned}\right\} \tag{13.48}$$

For Δ, the partial derivatives are obtained from the above equations by substituting Δ for Λ, and C_D for C_L.

For N:

$$\left.\begin{aligned}
\frac{\partial N}{\partial r} &= N\left\{\frac{1}{\rho}\frac{d\rho}{dr}\left(1+\frac{\mathrm{Re}}{C_M}\frac{\partial C_M}{\partial \mathrm{Re}}\right)+\frac{1}{V^2}\frac{dw}{dr}\left[\left(\frac{M}{C_M}\frac{\partial C_M}{\partial M}+\frac{\mathrm{Re}}{C_M}\frac{\partial C_M}{\partial \mathrm{Re}}+2\right)\right.\right.\\
&\qquad\left.\left.\cdot(v_\theta+w)+\frac{1}{C_M}\frac{\partial C_M}{\partial \alpha}v_r\right]-\frac{M}{C_M}\frac{\partial C_M}{\partial M}\frac{1}{a}\frac{da}{dr}-\frac{\mathrm{Re}}{C_M}\frac{\partial C_M}{\partial \mathrm{Re}}\frac{1}{\mu}\frac{d\mu}{dr}\right\}\\
\frac{\partial N}{\partial v_r} &= N\frac{v_r}{V^2}\left(\frac{M}{C_M}\frac{\partial C_M}{\partial M}+\frac{\mathrm{Re}}{C_M}\frac{\partial C_M}{\partial \mathrm{Re}}+2-\frac{1}{C_M}\frac{\partial C_M}{\partial \alpha}\frac{v_\theta+w}{v_r}\right)\\
\frac{\partial N}{\partial v_\theta} &= N\frac{v_\theta+w}{V^2}\left(\frac{M}{C_M}\frac{\partial C_M}{\partial M}+\frac{\mathrm{Re}}{C_M}\frac{\partial C_M}{\partial \mathrm{Re}}+2+\frac{1}{C_M}\frac{\partial C_M}{\partial \alpha}\frac{v_r}{v_\theta+w}\right)\\
\frac{\partial N}{\partial \beta} &= N\frac{1}{C_M}\frac{\partial C_M}{\partial \alpha}\\
\frac{\partial N}{\partial \gamma} &= N\frac{1}{C_M}\frac{\partial C_M}{\partial \gamma}\\
\frac{\partial N}{\partial \rho} &= N\frac{1}{\rho}\left(1+\frac{\mathrm{Re}}{C_M}\frac{\partial C_M}{\partial \mathrm{Re}}\right)\\
\frac{\partial N}{\partial T} &= -N\left(\frac{M}{C_M}\frac{\partial C_M}{\partial M}\frac{1}{2T}+\frac{\mathrm{Re}}{C_M}\frac{\partial C_M}{\partial \mathrm{Re}}\frac{1}{\mu}\frac{\partial \mu}{\partial T}\right)\\
\frac{\partial N}{\partial w} &= \frac{\partial N}{\partial v_\theta}\\
\frac{\partial N}{\partial I} &= -\frac{N}{I}
\end{aligned}\right\} \quad (13.49)$$

With these partial derivatives, the coefficient a's, b's, and c's can be easily calculated:

$$\left.\begin{aligned}
a_1 &= \frac{\partial F}{\partial r} = \frac{\partial \Sigma}{\partial r}\sin\beta+\frac{dw}{dr}\Lambda+(v_\theta+w)\frac{\partial \Lambda}{\partial r}-v_r\frac{\partial \Delta}{\partial r}+\left(\frac{v_\theta}{r}\pm\Omega\right)^2\\
&\qquad -2\frac{v_\theta}{r}\left(\frac{v_\theta}{r}\pm\Omega\right)+2\frac{g}{r}\left(\frac{r_0}{r}\right)^2\\
a_2 &= \frac{\partial F}{\partial \beta} = \Sigma\cos\beta+(v_\theta+w)\frac{\partial \Lambda}{\partial \beta}-v_r\frac{\partial \Delta}{\partial \beta}\\
a_3 &= \frac{\partial F}{\partial v_r} = (v_\theta+w)\frac{\partial \Lambda}{\partial v_r}-\Delta-v_r\frac{\partial \Delta}{\partial v_r}\\
a_4 &= \frac{\partial F}{\partial v_\theta} = \Lambda+(v_\theta+w)\frac{\partial \Lambda}{\partial v_\theta}-v_r\frac{\partial \Delta}{\partial v_\theta}+2\left(\frac{v_\theta}{r}\pm\Omega\right)\\
a_5 &= \frac{\partial F}{\partial \gamma} = (v_\theta+w)\frac{\partial \Lambda}{\partial \gamma}-v_r\frac{\partial \Delta}{\partial \gamma}\\
a_6 &= \frac{\partial F}{\partial \rho} = (v_\theta+w)\frac{\partial \Lambda}{\partial \rho}-v_r\frac{\partial \Delta}{\partial \rho}\\
a_7 &= \frac{\partial F}{\partial T} = (v_\theta+w)\frac{\partial \Lambda}{\partial T}-v_r\frac{\partial \Delta}{\partial T}\\
a_8 &= \frac{\partial F}{\partial w} = \Lambda+(v_\theta+w)\frac{\partial \Lambda}{\partial w}-v_r\frac{\partial \Delta}{\partial w}\\
a_9 &= \frac{\partial F}{\partial W} = \frac{\partial \Sigma}{\partial W}\sin\beta+(v_\theta+w)\frac{\partial \Lambda}{\partial W}-v_r\frac{\partial \Delta}{\partial W}
\end{aligned}\right\} \quad (13.50)$$

$$
\left.
\begin{aligned}
b_1 &= \frac{\partial G}{\partial r} = \frac{\partial \Sigma}{\partial r}\cos\beta - v_r\frac{\partial \Lambda}{\partial r} - \frac{dw}{dr}\Delta - (v_\theta + w)\frac{\partial \Delta}{\partial r} + \frac{v_r v_\theta}{r^2} \\
b_2 &= \frac{\partial G}{\partial \beta} = -\Sigma\sin\beta - v_r\frac{\partial \Lambda}{\partial \beta} - (v_\theta + w)\frac{\partial \Delta}{\partial \beta} \\
b_3 &= \frac{\partial G}{\partial v_r} = -\Lambda - v_r\frac{\partial \Lambda}{\partial v_r} - (v_\theta + w)\frac{\partial \Delta}{\partial v_r} - 2\left(\frac{1}{2}\frac{v_\theta}{r} \pm \Omega\right) \\
b_4 &= \frac{\partial G}{\partial v_\theta} = -v_r\frac{\partial \Lambda}{\partial v_\theta} - \Delta - (v_\theta + w)\frac{\partial \Delta}{\partial v_\theta} - \frac{v_r}{r} \\
b_5 &= \frac{\partial G}{\partial \gamma} = -v_r\frac{\partial \Lambda}{\partial \gamma} - (v_\theta + w)\frac{\partial \Delta}{\partial \gamma} \\
b_6 &= \frac{\partial G}{\partial \rho} = -v_r\frac{\partial \Lambda}{\partial \rho} - (v_\theta + w)\frac{\partial \Delta}{\partial \rho} \\
b_7 &= \frac{\partial G}{\partial T} = -v_r\frac{\partial \Lambda}{\partial T} - (v_\theta + w)\frac{\partial \Delta}{\partial T} \\
b_8 &= \frac{\partial G}{\partial w} = -v_r\frac{\partial \Lambda}{\partial w} - \Delta - (v_\theta + w)\frac{\partial \Delta}{\partial w} \\
b_9 &= \frac{\partial G}{\partial W} = \frac{\partial \Sigma}{\partial W}\cos\beta - v_r\frac{\partial \Lambda}{\partial W} - (v_\theta + w)\frac{\partial \Delta}{\partial W}
\end{aligned}
\right\}
\tag{13.51}
$$

$$
\left.
\begin{aligned}
c_1 &= \frac{\partial H}{\partial r} = -\frac{1}{r^2}\left[\Sigma\cos\beta - v_r\Lambda - (v_\theta + w)\Delta - 2v_r\left(\frac{v_\theta}{r} \pm \Omega\right)\right] \\
&\qquad + \frac{1}{r}\left[\frac{\partial \Sigma}{\partial r}\cos\beta - v_r\frac{\partial \Lambda}{\partial r} - (v_\theta + w)\frac{\partial \Delta}{\partial r} - \frac{dw}{dr}\Delta + 2\frac{v_r v_\theta}{r^2}\right] + \frac{\partial N}{\partial r} \\
c_2 &= \frac{\partial H}{\partial \beta} = \frac{1}{r}\left[-\Sigma\sin\beta - v_r\frac{\partial \Lambda}{\partial \beta} - (v_\theta + w)\frac{\partial \Delta}{\partial \beta}\right] + \frac{\partial N}{\partial \beta} \\
c_3 &= \frac{\partial H}{\partial v_r} = \frac{1}{r}\left[-\Lambda - v_r\frac{\partial \Lambda}{\partial v_r} - (v_\theta + w)\frac{\partial \Delta}{\partial v_r} - 2\left(\frac{v_\theta}{r} \pm \Omega\right)\right] + \frac{\partial N}{\partial v_r} \\
c_4 &= \frac{\partial H}{\partial v_\theta} = \frac{1}{r}\left[-v_r\frac{\partial \Lambda}{\partial v_\theta} - \Delta - (v_\theta + w)\frac{\partial \Delta}{\partial v_\theta} - 2\frac{v_r}{r}\right] + \frac{\partial N}{\partial v_\theta} \\
c_5 &= \frac{\partial H}{\partial \gamma} = \frac{1}{r}\left[-v_r\frac{\partial \Lambda}{\partial \gamma} - (v_\theta + w)\frac{\partial \Delta}{\partial \gamma}\right] + \frac{\partial N}{\partial \gamma} \\
c_6 &= \frac{\partial H}{\partial \rho} = \frac{1}{r}\left[-v_r\frac{\partial \Lambda}{\partial \rho} - (v_\theta + w)\frac{\partial \Delta}{\partial \rho}\right] + \frac{\partial N}{\partial \rho} \\
c_7 &= \frac{\partial H}{\partial T} = \frac{1}{r}\left[-v_r\frac{\partial \Lambda}{\partial T} - (v_\theta + w)\frac{\partial \Delta}{\partial T}\right] + \frac{\partial N}{\partial T} \\
c_8 &= \frac{\partial H}{\partial w} = \frac{1}{r}\left[-v_r\frac{\partial \Lambda}{\partial w} - \Delta - (v_\theta + w)\frac{\partial \Delta}{\partial w}\right] + \frac{\partial N}{\partial w} \\
c_9 &= \frac{\partial H}{\partial W} = \frac{1}{r}\left[\frac{\partial \Sigma}{\partial W}\cos\beta - v_r\frac{\partial \Lambda}{\partial W} - (v_\theta + w)\frac{\partial \Delta}{\partial W}\right] + \frac{\partial N}{\partial W} \\
c_{10} &= \frac{\partial H}{\partial I} = \frac{\partial N}{\partial I}
\end{aligned}
\right\}
\tag{13.52}
$$

After the power cutoff, the thrust S vanishes. Thus for $t > \bar{t}_1$, Σ and its derivatives are zero.

CHAPTER 14

CONTROL DESIGN WITH SPECIFIED CRITERIA

In the earlier chapters, we discussed the design problem of control systems mainly from the point of view of analysis. That is, assume the construction of the system, and then find out whether the performance of the system is satisfactory. In the last chapter, we introduced for the first time a different and more direct point of view: we specify the performance first, and then find out the necessary control system which will give the desired performance. In this chapter, we shall extend this principle to arbitrarily controlled systems, for which the performance criteria are expressed in terms of integrals of the controlled variables. The resultant system behavior is represented by a very general equation, which is usually nonlinear. Thus the control system so designed is generally a nonlinear system. The nonlinearity here, however, is purposefully chosen to give the optimum performance of the over-all system.

Mathematically, we can describe the principle of control design with specified criteria as follows. In the system to be controlled, we introduce one or more extra variables. These extra variables, being artificially created, are not determined by the intrinsic physical laws of the controlled system. We obtain the conditions for determining the extra variables by satisfying the specified criteria of the over-all performance. These conditions are then enforced through the computer built into the system. This principle of control design was first suggested by Boksenbom and Hood.[1] The following discussion follows partly the cited work of these authors.

14.1 Control Criteria. If y is the output of the controlled system, then it is reasonable to expect that the measure of over-all performance of the system is expressed as the time integral of some function f of y. Then the criterion of the performance is that this integral is to be minimum or a constant; that is,

$$\int_0^{t_1} f(y)\,dt = \text{const. or min.} \tag{14.1}$$

or, specifically,

$$\int_0^{t_1} (y - y_s)^2\,dt = \text{const. or min.} \tag{14.2}$$

[1] A. S. Boksenbom and R. Hood, *NACA TR* 1068 (1952).

where t_1 is the time at the end of transient and y_s is the setting or the desired value of y. Equation (14.2) weights the error in y as the square and according to the time duration of that error, and is thus a measure of the mean-square error from the setting. Another type of criterion may be that which requires a criterion time duration to be a minimum or a constant; that is,

$$\int_0^{t_1} dt = \text{const. or min.} \tag{14.3}$$

The use of a single criterion, such as Eq. (14.1), will yield $f(y)$ = constant. This result is reasonable because $f(y)$ can usually be made to be a constant if no additional criteria are imposed on other variables in the system. Usually, however, certain limiting conditions exist on other variables in the system, and these conditions must be included in the original criterion. Thus, for example, a possible criterion could be written as follows:

$$\left.\begin{aligned} \int_0^{t_1} (y - y_s)^2\, dt &= \text{min.} \\ \text{for} \qquad \int_0^{t_1} f(z)\, dt &= \text{const.} \end{aligned}\right\} \tag{14.4}$$

If, for instance, y is the engine speed and z is the characteristic temperature of a gas turbine engine, the criterion of (14.4) states that it is desirable to design a control system such that, for a particular value of a temperature integral, the integral of the speed-error squared is a minimum. This criterion may be used if, for instance, it is known that an overtemperature condition can be tolerated for a certain period of time and it is desired to keep the average speed error at a minimum during this transient. Then the integral of z represents the total heat input to the turbine blades.

The general theory will show that as many criteria as desired of the type shown in Eqs. (14.1) to (14.4) can be included together, and a control system can be derived that automatically satisfies all these criteria simultaneously.

Another aspect of the control criteria is the end conditions of the integrals of Eqs. (14.1) to (14.4). The time interval for which these integrals are to be a minimum or a constant must be chosen. A reasonable time interval is any duration during which essential external disturbances are constant and during which the system to be controlled moves from one essential level of operation to another. The essential external disturbances are those that cannot be immediately corrected by the control system. If an essential external disturbance occurs in the time interval of the criteria, no physically realizable system could be designed to anticipate this disturbance so as to behave properly before the disturbance takes place. An essential level of operation is any

specific condition of only those variables that must be continuous. In the case of a turbojet engine, the transient behavior of which can be described by a first-order differential equation, the engine speed determines the level of operation. If the fuel system must be considered or if the temperature does not respond to speed immediately, then both the engine and the acceleration are required to describe the essential level of the engine. We shall see this presently.

The control system resulting from any design method must be physically realizable. There are two aspects to this problem. First, it is possible to set down criteria that are not realizable with any system or are incompatible with each other. If such criteria are used, the unrealizability will appear either as a requirement on the control to look ahead into the future or as an inability to satisfy the boundary conditions of some differential equation. In most cases, a clear understanding of the criteria used and of the system to be controlled will preclude incompatibilities of this sort.

The second aspect of physical realizability is purely mathematical. It is desired to derive a description (a differential equation) of the control or the controlled system that satisfies the criteria of control and all the necessary boundary conditions that arise in the derivation of this equation. Although the mathematical solution of the problem may be any derivative or integral of this differential equation, the physical solution of the problem requires the differential equation that itself satisfies the boundary conditions and for which no undetermined constants of integration exist. Thus, such forms as

$$\dot{y} = c\dot{x}$$

and

$$y = cx$$

are not necessarily interchangeable as descriptions of some part of a controlled system, because the forms differ by an undetermined constant of integration. For stable linear systems, the effect of this constant becomes vanishingly small; for the general nonlinear systems presented here, however, this constant must be considered.

14.2 Stability Problem. The requirement of stability is a special criterion that does not enter into the design of the main control system during the transient. This situation is the same as that of the last chapter where the stability criterion is also suppressed, because the satisfactory performance of the over-all system is already made certain by the imposed performance specifications. However, it is usually necessary to add to the controlled system a stability device that does not go into action until the final instant of the transient interval. Therefore, this stability device will not affect the behavior of the system as

far as satisfying the other criteria is concerned. This device can be described as follows: For a first-order system, when

$$\left.\begin{aligned} y &= y_s \\ \text{then} \qquad \dot{y} &= 0 \end{aligned}\right\} \tag{14.5}$$

For second-order systems, when

$$\left.\begin{aligned} &\qquad y = y_s \\ \text{then} \qquad \dot{y} = 0 \quad &\text{and} \quad \ddot{y} = 0 \end{aligned}\right\} \tag{14.6}$$

When such a stability device is added to the control system, the control system has two modes of operation and is thus a *multiple-mode system* (*cf.* Sec. 10.9). During the transient, the main control system is in operation to produce the specified performance. At the end of transient, the control is switched to a second system, represented by Eq. (14.5) or Eq. (14.6), to ensure stability of the system at the end state and thus to avoid running away from the desired operating point of the system.

14.3 General Theory for First-order Systems. With the type of control criteria given previously by Eq. (14.1) to (14.4), we can formulate the control equation in the following manner: if, for such a list of criteria, one of the integrals is to be a minimum under the condition that the other integrals are to be constant, it is sufficient, according to variational calculus, to make

$$\int_0^{t_1} f(y)\,dt + \lambda_1 \int_0^{t_1} (y - y_s)^2\,dt + \lambda_2 \int_0^{t_1} f_0(z)\,dt + \lambda_3 \int_0^{t_1} dt = \text{min.}$$

or

$$\int_0^{t_1} [f(y) + \lambda_1(y - y_s)^2 + \lambda_2 f_0(z) + \lambda_3]\,dt = \text{min.} \tag{14.7}$$

The λ's are arbitrary constants that enter into the control system as the adjustable parameters and are precisely determined by the choice of values that the constant integrals are to have. The technique of the λ multipliers is widely used for problems of this type, where one condition is to be a minimum under other restrictive conditions. Indeed, the conditions need not be in integral form, and any functional or differential relation among variables can be handled in a similar manner.[1] Equation (14.7) can be made very general when all possible restrictive conditions are included. In the final equations, if any one criterion is not to be used, then the corresponding $\lambda \to 0$. If any of the criteria are to be zero, then the corresponding $\lambda \to \infty$.

[1] See for instance C. Lanczos, "The Variational Principles of Mechanics," University of Toronto Press, Toronto, 1946.

If the system to be controlled is of *first order* with constant coefficients and with one essential output y, then the variables y and z must be related so that $z = z(y,\dot{y})$. Equation (14.7) can then be written, in general, as

$$\int_0^{t_1} F(y,\dot{y})\, dt = \min. \tag{14.8}$$

where F is a continuous function of y and $\dot{y}$, and y is a continuous function of time t We note that F is not explicitly dependent on the time t.

Let us consider $y(t)$ to be a solution, that is, $y(t)$ is the output among all admissible outputs which satisfies the condition of Eq. (14.8). To

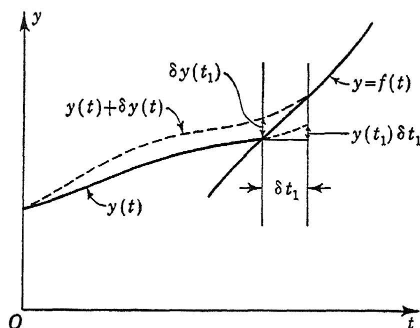


FIG. 14.1

test this, we construct a neighboring solution as $y(t) + \epsilon\, \delta y(t)$. $\delta y(t)$ is an arbitrary function, and ϵ is a small parameter. If the condition of Eq. (14.8) is satisfied by $y(t)$, then

$$\left[\frac{d}{d\epsilon}\int_0^{t_1+\epsilon\,\delta t_1} F(y + \epsilon\,\delta y, \dot{y} + \epsilon\,\delta\dot{y})\, dt\right]_{\epsilon=0} = 0$$

or

$$\int_0^{t_1} \frac{\partial F}{\partial y}\,\delta y\, dt + \int_0^{t_1} \frac{\partial F}{\partial \dot{y}}\,\delta\dot{y}\, dt + F(t_1)\,\delta t_1 = 0 \tag{14.9}$$

The variation δt_1 occurs because of the fact that the end point of the integral of Eq. (14.8) is not fixed, but lies on the curve $y = f(t)$, as shown in Fig. 14.1. This is to allow the proper boundary conditions of moving from one essential level of operation to another, as previously discussed. By partial integration, Eq. (14.9) becomes

$$\int_0^{t_1}\left[\frac{\partial F}{\partial y} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right)\right]\delta y\, dt + \left(\frac{\partial F}{\partial \dot{y}}\right)_{t_1}\delta y(t_1) - \frac{\partial F}{\partial \dot{y}}\,\delta y(0) + F(t_1)\,\delta t_1 = 0$$

The relationship between $\delta y(t_1)$ and δt_1 can be easily computed from the end condition. That is,

$$\dot{y}\,\delta t_1 + \delta y(t_1) = f'(t_1)\,\delta t_1$$

Then by eliminating $\delta y(t_1)$, and since δt is arbitrary, we have

$$\int_0^{t_1} \left[\frac{\partial F}{\partial y} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right)\right] \delta y \, dt = 0 \qquad (14.10)$$

and

$$\delta t_1 \left\{F(t_1) + \left(\frac{\partial F}{\partial \dot{y}}\right)_{t_1} \left[f'(t_1) - \dot{y}(t_1)\right]\right\} - \left(\frac{\partial F}{\partial \dot{y}}\right)_0 \delta y(0) = 0 \qquad (14.11)$$

The time interval during which the criterion of Eq. (14.8) is to hold is considered as that during which the system moves from one essential operating level to another; in this case, from one definite value of y to another definite value of y. Thus the end curve $y = f(t)$ must be a straight line with $f(t) =$ constant. Hence

$$\left.\begin{aligned} \delta y(0) &= 0 \\ f'(t_1) &= 0 \end{aligned}\right\} \qquad (14.12)$$

Thus Eqs. (14.10) and (14.11) become

$$\frac{\partial F}{\partial y} = \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right) \qquad (14.13)$$

and

$$F(t_1) = \dot{y}(t_1)\left(\frac{\partial F}{\partial \dot{y}}\right)_{t_1} \qquad \text{if } \left(\frac{\partial F}{\partial \dot{y}}\right)_0 \text{ is finite} \qquad (14.14)$$

Equation (14.13) need not hold at $t = 0$ because $\delta y(0) = 0$. The only conditions that need hold at $t = 0$ are that $(\partial F/\partial \dot{y})_0$ is finite and y is continuous. At the start of a new transient, $\dot{y}$, F, $(\partial F/\partial y)$, and $(\partial F/\partial \dot{y})$ may be discontinuous, whereas at other points $(0 < t \leq t_1)$, $\partial F/\partial \dot{y}$ will be continuous because of Eq. (14.13).

Equation (14.13) is the differential equation for the $y(t)$ that satisfies the original criterion of Eq. (14.8). This equation is the so-called *Euler-Lagrange differential equation* of our variational problem. For the problem considered here, F does not explicitly depend upon t. Then we can immediately obtain a first integral of this equation. The first integral of Eq. (14.13), which satisfies the boundary condition of Eq. (14.14), is

$$F(y,\dot{y}) = \dot{y}\,\frac{\partial F}{\partial \dot{y}} \qquad (14.15)$$

By differentiating this equation with respect to t, we have

$$\frac{\partial F}{\partial y}\,\dot{y} + \ddot{y}\,\frac{\partial F}{\partial \dot{y}} = \ddot{y}\,\frac{\partial F}{\partial \dot{y}} + \dot{y}\,\frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right)$$

or whenever y, $\partial F/\partial \dot{y}$, and so forth are continuous

$$\dot{y}\left[\frac{\partial F}{\partial y} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right)\right] = 0$$

Thus either $\dot{y} = 0$ or Eq. (14.13) is satisfied. However, $\dot{y}$ does not generally vanish during the transient. Thus the two conditions on $y(t)$ as specified by Eqs. (14.13) and (14.14) are now replaced by the single equation (14.15).

Thus Eq. (14.15) is the description of that physically realizable system the behavior of which will automatically and simultaneously satisfy those criteria included in the function F during that time interval for which the external disturbances are constant and during which the system goes from one operating level to any other operating level. At the end of the transient when the end level of y is reached, a stability device must be added to the system; the description of such an ideal device is that of Eq. (14.5).

14.4 Application to Turbojet Controls. In the usual case of designing turbojet engine controls, the engine speed N that sets the essential operating level of the engine is to be set or controlled. As the result, other pertinent characteristics, such as thrust, are also set. Limiting conditions of the engine are those of overspeed, overtemperature, compressor surge, and rich burner blowout. Let N_s be the engine speed setting, T be the inlet temperature to the turbine, and P be the discharge pressure of the compressor; then the criteria on the behavior of this engine can be specified as the following integrals:

$$\left.\begin{array}{ll}
\int_0^{t_1} f_1(N - N_s)\,dt & \text{for speed control} \\
\int_0^{t_1} f_2(N)\,dt & \text{for speed overshoot} \\
\int_0^{t_1} f_3(T)\,dt & \text{for temperature overshoot and undershoot} \\
\int_0^{t_1} f_4[P - g(N)]\,dt & \text{for compressor surge} \\
\int_0^{t_1} f_5[P - h(N)]\,dt & \text{for blowout} \\
\text{and} \quad \int_0^{t_1} dt & \text{for rise time}
\end{array}\right\} \tag{14.16}$$

The nature of these functions is sketched in Fig. 14.2. The quantity $P - g(N)$ is the amount by which the compressor discharge pressure exceeds the safe pressure for surge, and $g(N)$ is the compressor discharge pressure for each engine speed at a safe value below surge. Rich burner blowout can be handled in a similar manner. The rise time is the total time for the system to move from one essential operating level to the other.

Similarly to the treatment on turbojet behavior in Sec. 5.6, the linearized engine characteristics can be expressed as follows:

$$\left.\begin{array}{l}
T = aN + a_T\dot{N} \\
P = bN + c\dot{T}
\end{array}\right\} \tag{14.17}$$

where τ is the engine time constant. By substituting these relations into the integrals of Eq. (14.16), we see that they all take the form

$$\int_0^{t_1} f(N,\dot{N})\, dt$$

where f is a continuous function of N and $\dot{N}$, and N is a continuous function of t.

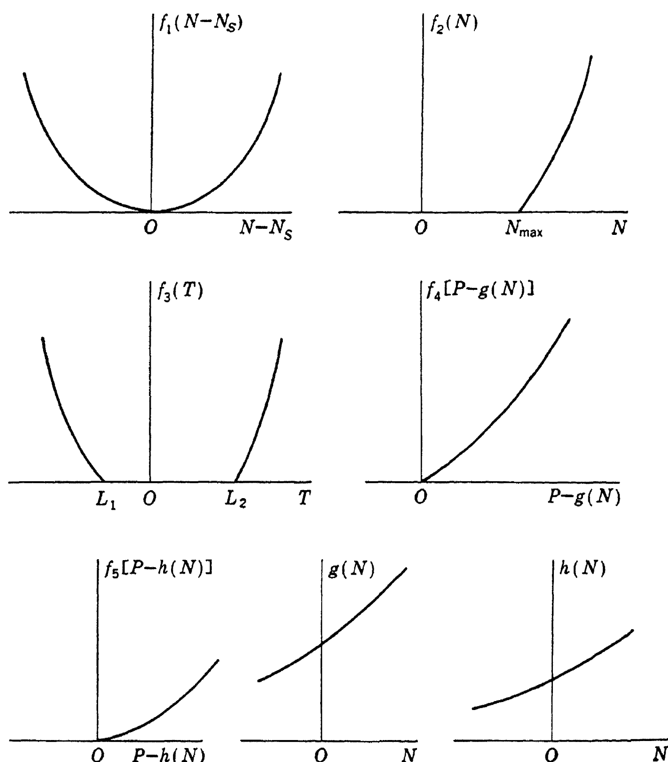


Fig. 14.2

14.5 Speed Control with Temperature-limiting Criteria. If only the error in speed control is considered important, the criterion becomes

$$\int_0^{t_1} f(N - N_s)\, dt = \text{min.}$$

Then the control condition of Eq. (14.15) simply gives

$$f_1(N - N_s) = 0$$

Then because of the nature of the function f_1, $N = N_s$. This result means that, in the absence of other criteria on the engine behavior, this speed control should keep the speed error identically zero, which is

physically possible only in the sense of allowing infinite temperatures. This result, however, is inconsistent with the previous development of Eq. (14.15), in that N is not a discontinuous function of time. This instance is actually a trivial case of the general problem. But the result does indicate that a criterion like this must be accompanied by an additional criterion to give a physically sensible system.

Now if the error in speed control is to be combined with the condition on the overshoot and undershoot of temperature, then

$$\int_0^{t_1} [f_1(N - N_s) + \lambda f_3(T)]\, dt = \min. \tag{14.18}$$

Therefore $F = f_1(N - N_s) + \lambda f_3(T)$. By using Eq. (14.17), Eq. (14.15) becomes

$$f_1(N - N_s) + \lambda f_3(T) = \lambda a \tau \dot{N} f_3'(T) \tag{14.19}$$

This is the control equation during the transient. At the end of the transient, the ideal stability device is switched on, so that when

$$\left.\begin{aligned} N &= N_s \\ \text{then} \qquad N &= 0 \end{aligned}\right\} \tag{14.20}$$

Equations (14.19) and (14.20) describe the complete control system. Therefore we can visualize a computer so designed that it takes information from the measurements on N and T, the stored information on λ, a, and τ, and the relation between the fuel rate and N and T, and generates the signal for the proper fuel rate in accordance with Eq. (14.19). Then shortly before N reaches N_s, the stability device takes over, so that at the end of the transient, Eq. (14.20) is satisfied. In general, the control equation (14.19) is nonlinear, and the computer cannot be a linear device, such as a simple RC circuit.

As an example it is convenient to let $f_3(T) = (T - L_2)^n$ for $T > L_2$ and $f_3(T) = (L_1 - T)^n$ for $T < L_1$. In general, the power n should be > 1, because when $n < 1$, T may be infinite and of such nature as to make N discontinuous and physically unreal, even though the integral

$$\int_0^{t_1} f_3(T)\, dt$$

is finite. In the example of this discussion, let $n = 2$, and, furthermore, $f_1(N - N_s) = (N - N_s)^2$. Therefore we again take the mean-square deviation from the setting as a measure of the error. Then Eq. (14.19) becomes

$$\frac{(N - N_s)^2}{\lambda} + (L - aN)^2 = a^2 \tau^2 \dot{N}^2 \tag{14.21}$$

where, for acceleration, or when $N < N_s$,

$$\dot{N} > 0 \qquad \text{and} \qquad L = L_2 \tag{14.22}$$

For deceleration, or $N > N_s$,

$$\dot{N} < 0 \quad \text{and} \quad L = L_1 \tag{14.23}$$

The block diagram for the control system is shown in Fig. 14.3. N_e is the actual engine speed. We assume the case of deceleration with $N > N_s$. During the transient, $N_e - N_s$ is positive, the switch between the computer and the engine servomechanism is closed, and the signal from the computer is in action. The computer generates the signal $a\tau\dot{N}$ according to the control Eq. (14.21). In Fig. 14.3, the signal is schematically indicated by a rectangular triangle. The engine servomechanism is so designed that the signal $a\tau\dot{N}$ from the computer is actually obeyed closely by the engine. This is done by using a high-

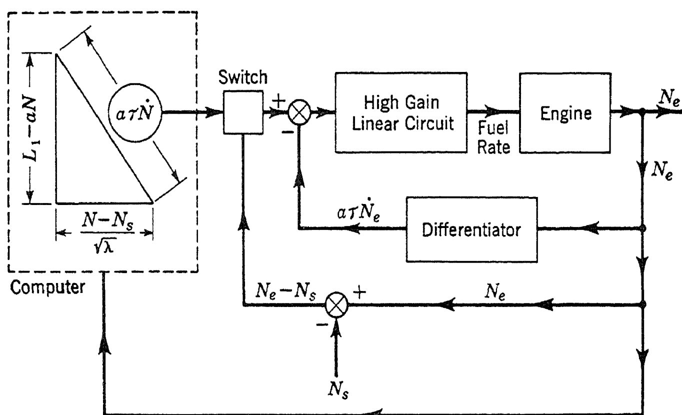


FIG. 14.3

gain circuit as indicated. When the speed error $(N_e - N_s)$ is reduced to a very small value, the computer signal is switched off. Then the stability of the engine servomechanism will guarantee the stability of the system at the engine setting N_s, and the system essentially satisfies the condition of Eq. (14.20).

The control system has one adjustable parameter λ. For any value of λ, this system will, for the value of integral temperature overshoot obtained, give the minimum value of the integral speed error squared. The value of λ determines the actual value of the integral temperature overshoot. Let us consider *the special case* where $aN_s = L$; that is, acceleration or deceleration to the speed that corresponds to the limiting temperatures occurs, according to Eq. (14.17). It is interesting to note that for this special case, $\dot{N} = 0$ when $N = N_s$, according to the control condition of Eq. (14.21). Therefore no separate stability device is necessary, and the switch in the control system as shown by Fig. 14.3 can be

eliminated. In this special case, Eq. (14.21) becomes linear and can be written as

$$E(L - aN) = a\tau\dot{N} \tag{14.24}$$

where

$$E = \left(1 + \frac{1}{a^2\lambda}\right)^{\frac{1}{2}} \tag{14.25}$$

Now the integrals can be easily computed. For instance, the temperature integral is

$$\begin{aligned}
\int_0^{t_1} (T - L)^2\, dt &= \int_0^{t_1} (aN - L + a\tau\dot{N})^2\, dt \\
&= (E - 1)^2 \int_0^{t_1} (L - aN)^2\, dt \\
&= a^2(E - 1)^2 \int_0^{t_1} (N_s - N)^2\, dt \\
&= a^2(E - 1)^2 \int_{N_0}^{N_s} (N_s - N)^2 \frac{dN}{\dot{N}} \\
&= a^3\tau(E - 1)^2 \int_{N_0}^{N_s} \frac{(N_s - N)^2\, dN}{Ea(N_s - N)} \\
&= a^2\tau \frac{(E - 1)^2}{E} \frac{1}{2} (N_s - N_0)^2 = \frac{1}{2} \frac{(E - 1)^2}{E} (L - aN_0)^2
\end{aligned}$$

Thus

$$\frac{1}{\tau} \int_0^{t_1} \frac{(T - L)^2}{(L - aN_0)^2}\, dt = \frac{(E - 1)^2}{2E} \tag{14.26}$$

where N_0 is the engine speed at the beginning of the transient. Similarly, the speed integral is

$$\frac{a^2}{\tau} \int_0^{t_1} \left(\frac{N - N_s}{L - aN_0}\right)^2 dt = \frac{1}{2E} \tag{14.27}$$

and if T_{max} is the maximum temperature, then

$$\frac{T_{max} - L}{L - aN_0} = E - 1 \tag{14.28}$$

From Eq. (14.24), we have

$$Ea(N_s - N) = a\tau \frac{dN}{dt}$$

so that the characteristic time τ^* for the controlled transient is

$$\tau^* = \frac{\tau}{E} \tag{14.29}$$

The left sides of these equations have been put in dimensionless form. The maximum temperature T_{max} occurs at the beginning of the transient.

For $E = 1$ ($\lambda = \infty$), the temperature does not overshoot, in agreement with our previous statement that the constant integral has the value zero when $\lambda \to \infty$. The speed integral is 0.5, and $\tau^* = \tau$. As E increases (or λ decreases), the temperature integral and the maximum temperature increase, whereas the speed integral and the time constant decrease. A compromise value for E may be $\sqrt{2}$, or $a^2\lambda = 1$. Once the choice of E or λ is settled, the specification for the control computer is fixed. The design can then proceed.

For the general case of Eq. (14.21), the calculation of the values of the integrals is somewhat more cumbersome, but a similar procedure applies for the design of the control system. Boksenbom and Hood actually give solutions of Eq. (14.26), that is, N as a function of time t. But we should emphasize here that such explicit solutions are not necessary for the control design. The information for control design is fully set forth by Eq. (14.24) itself, because that equation tells how the control computer should be constructed. If the control computer is made according to that condition, then the desired performance is ensured. The actual variation of N with respect to time is thus of no importance. Hence our approach to the design problem is to "design" the nonlinear control equation itself rather than to design according to the solutions of an assumed equation.

14.6 Second-order Systems with Two Degrees of Freedom. For the case of a second-order system with two degrees of freedom and with constant coefficients, Eq. (14.8) becomes

$$\int_0^{t_1} F(y,\dot{y},\ddot{y},z,\dot{z},\ddot{z})\, dt = \min. \tag{14.30}$$

where y and z are the outputs and are independent functions of time. The condition to satisfy Eq. (14.30) is

$$\left.\begin{aligned} \frac{d}{d\epsilon}\int_0^{t_1+\epsilon\,\delta t_1} F(y + \epsilon\,\delta y,\dot{y} + \epsilon\,\delta\dot{y},\ddot{y} + \epsilon\,\delta\ddot{y},\\ z + \epsilon\,\delta z,\dot{z} + \epsilon\,\delta\dot{z},\ddot{z} + \epsilon\,\delta\ddot{z})\, dt = 0\\ \text{at } \epsilon = 0 \end{aligned}\right\} \tag{14.31}$$

The time interval of the integral of Eq. (14.30) begins at a definite time ($t = 0$) but does not end at any definite time, but rather along the curves $y = f_1(t)$, $\dot{y} = f_2(t)$, $z = g_1(t)$, and $\dot{z} = g_2(t)$. The functions δy and δz are arbitrary and, naturally, independent functions of time.

Performing the operation indicated by Eq. (14.31) gives

$$\int_0^{t_1}\left[\frac{\partial F}{\partial y}\,\delta y + \frac{\partial F}{\partial \dot{y}}\,\delta\dot{y} + \frac{\partial F}{\partial \ddot{y}}\,\delta\ddot{y} + \frac{\partial F}{\partial z}\,\delta z + \frac{\partial F}{\partial \dot{z}}\,\delta\dot{z} + \frac{\partial F}{\partial \ddot{z}}\,\delta\ddot{z}\right] dt + F(t_1)\,\delta t_1 = 0$$

After integration by parts, we have

$$\int_0^{t_1}\left[\frac{\partial F}{\partial y}-\frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right)+\frac{d^2}{dt^2}\left(\frac{\partial F}{\partial \ddot{y}}\right)\right]\delta y\,dt$$
$$+\int_0^{t_1}\left[\frac{\partial F}{\partial z}-\frac{d}{dt}\left(\frac{\partial F}{\partial \dot{z}}\right)+\frac{d^2}{dt^2}\left(\frac{\partial F}{\partial \ddot{z}}\right)\right]\delta z\,dt$$
$$+\left[\frac{\partial F}{\partial \dot{y}}\,\delta y+\frac{\partial F}{\partial \ddot{y}}\,\delta \dot{y}-\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{y}}\right)\delta y\right]_0^{t_1}+F(t_1)\,\delta t_1$$
$$+\left[\frac{\partial F}{\partial \dot{z}}\,\delta z+\frac{\partial F}{\partial \ddot{z}}\,\delta \dot{z}-\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{z}}\right)\delta z\right]_0^{t_1}=0 \qquad (14.32)$$

As before, the integrands of the integrals and the boundary-condition terms must vanish separately. From the specified end condition, we have

$$\left.\begin{aligned}\delta y(t_1)&=[f_1'(t_1)-\dot{y}(t_1)]\,\delta t_1\\ \delta \dot{y}(t_1)&=[f_2'(t_1)-\ddot{y}(t_1)]\,\delta t_1\\ \delta z(t_1)&=[g_1'(t_1)-\dot{z}(t_1)]\,\delta t_1\\ \delta \dot{z}(t_1)&=[g_2'(t_1)-\ddot{z}(t_1)]\,\delta t_1\end{aligned}\right\} \qquad (14.33)$$

The three conditions from Eq. (14.32) are the two simultaneous Euler-Lagrange equations

$$\left.\begin{aligned}\frac{\partial F}{\partial y}-\frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right)+\frac{d^2}{dt^2}\left(\frac{\partial F}{\partial \ddot{y}}\right)&=0\\ \frac{\partial F}{\partial z}-\frac{d}{dt}\left(\frac{\partial F}{\partial \dot{z}}\right)+\frac{d^2}{dt^2}\left(\frac{\partial F}{\partial \ddot{z}}\right)&=0\end{aligned}\right\} \qquad (14.34)$$

and

$$\left\{\left[F-\dot{y}\frac{\partial F}{\partial \dot{y}}+\dot{y}\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{y}}\right)-\ddot{y}\frac{\partial F}{\partial \ddot{y}}-\dot{z}\frac{\partial F}{\partial \dot{z}}+\dot{z}\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{z}}\right)-\ddot{z}\frac{\partial F}{\partial \ddot{z}}\right]_{t=t_1}\right.$$
$$+f_1'(t_1)\left[\frac{\partial F}{\partial \dot{y}}-\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{y}}\right)\right]_{t=t_1}+f_2'(t_1)\left(\frac{\partial F}{\partial \ddot{y}}\right)_{t=t_1}+g_1'(t_1)\left[\frac{\partial F}{\partial \dot{z}}-\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{z}}\right)\right]_{t=t_1}$$
$$\left.+g_2'(t_1)\left(\frac{\partial F}{\partial \ddot{z}}\right)_{t=t_1}\right\}\delta t_1+\delta y(0)\left[\frac{\partial F}{\partial \dot{y}}-\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{y}}\right)\right]_{t=0}+\delta \dot{y}(0)\left(\frac{\partial F}{\partial \ddot{y}}\right)_{t=0}$$
$$+\delta z(0)\left[\frac{\partial F}{\partial \dot{z}}-\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{z}}\right)\right]_{t=0}+\delta \dot{z}(0)\left(\frac{\partial F}{\partial \ddot{z}}\right)_{t=0}=0 \qquad (14.35)$$

Equation (14.34) is a system of two equations that satisfies the original criterion of Eq. (14.30). The physical solution to the problem must satisfy Eq. (14.34) and, in addition, satisfy the boundary-condition equations of Eq. (14.35). However since F does not explicitly depend upon t, a modification of this set of conditions is possible: If the first of Eqs. (14.34) is multiplied by $\dot{y}$ and the second by $\dot{z}$ and the equations are added, an exact derivative is formed the integral of which is as follows:

$$F - \dot{y}\frac{\partial F}{\partial \dot{y}} + \dot{y}\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{y}}\right) - \ddot{y}\frac{\partial F}{\partial \ddot{y}} - \dot{z}\frac{\partial F}{\partial \dot{z}} + \dot{z}\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{z}}\right) - \ddot{z}\frac{\partial F}{\partial \ddot{z}} = C \quad (14.36)$$

Since F is a function of $\dot{y}$ and $\ddot{y}$, it is unlikely that $\partial F/\partial \dot{y}$ and $\partial F/\partial \ddot{y}$ can be zero, or that $\partial F/\partial \ddot{y}$ can be a constant with respect to time. Then $\dfrac{\partial F}{\partial y} - \dfrac{d}{dt}\left(\dfrac{\partial F}{\partial \ddot{y}}\right)$ is not zero. Therefore it is reasonable to have

$$\delta y(0) = \delta \dot{y}(0) = f_1'(t_1) = f_2'(t_1) = 0$$

in Eq. (14.35). A similar argument applies to the variable z. Thus a reasonable set of boundary conditions is as follows:

$$\left.\begin{aligned} \delta y(0) &= 0 \qquad & f_1'(t_1) &= 0 \\ \delta \dot{y}(0) &= 0 & f_2'(t_1) &= 0 \\ \delta z(0) &= 0 & g_1'(t_1) &= 0 \\ \delta \dot{z}(0) &= 0 & g_2'(t_1) &= 0 \end{aligned}\right\} \quad (14.37)$$

These boundary conditions are also the ones corresponding to fixed starting values and fixed end values of y, $\dot{y}$, z, and $\dot{z}$, but with a variable interval t_1 for the transient. With the conditions of Eq. (14.37), Eq. (14.35) shows that the constant C to the right in Eq. (14.36) must vanish. The final solution to the problem of Eq. (14.30) is as follows:

$$F - \dot{y}\frac{\partial F}{\partial \dot{y}} + \dot{y}\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{y}}\right) - \ddot{y}\frac{\partial F}{\partial \ddot{y}} - \dot{z}\frac{\partial F}{\partial \dot{z}} + \dot{z}\frac{d}{dt}\left(\frac{\partial F}{\partial \ddot{z}}\right) - \ddot{z}\frac{\partial F}{\partial \ddot{z}} = 0 \quad (14.38)$$

and either one of the following equations:

$$\left.\begin{aligned} \frac{\partial F}{\partial y} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right) + \frac{d^2}{dt^2}\left(\frac{\partial F}{\partial \ddot{y}}\right) &= 0 \\ \frac{\partial F}{\partial z} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{z}}\right) + \frac{d^2}{dt^2}\left(\frac{\partial F}{\partial \ddot{z}}\right) &= 0 \end{aligned}\right\} \quad (14.39)$$

Equations (14.38) and (14.39) constitute a system of two equations for two unknowns y and z. They are the control equations and are the equations for the computer design.

The boundary condition (14.37) defines the original criteria for that duration during which the system moves from one essential operating level to another. Thus, if all conditions of Eq. (14.37) must hold, the system goes from one definite y, $\dot{y}$, z, and $\dot{z}$ to any other definite y, $\dot{y}$, z, and $\dot{z}$. Equation (14.38) is a third-order equation. Equation (14.39) is a fourth-order equation. Thus besides the four initial values y, $\dot{y}$, z, $\dot{z}$ we can still assign three values of $\dot{y}$, z, $\dot{z}$ at the final y. That is, when $y = y_s$; $\dot{y} = 0$, $z = z_s$, and $\dot{z} = 0$. A stability device must still be added to the system, so that

$$\ddot{y} = 0 \qquad \text{and} \qquad \ddot{z} = 0$$

at the final point. We see then that although the system considered above is considerably more involved than the first-order system discussed in the previous sections, yet the same general approach is entirely applicable.

14.7 Control Problem with Differential Equation as Auxiliary Condition. Let us consider y to be the essential variable whose performance has to be controlled. z is the variable which we put in the system to ensure that y can be made to have the desired performance. The inherent dynamics of the system then gives one a relation between y and z. This relation is in general a nonlinear differential equation, say of second order,

$$g(y,\dot{y},\ddot{y};z,\dot{z},\ddot{z};t) = 0 \tag{14.40}$$

The performance of y during a transient is specified, say, as

$$\int_0^{t_1} f(y)\, dt = \text{min.} \tag{14.41}$$

The error integral of Eq. (14.2), would be an example of this type of specification. The problem is to derive a control equation satisfying both Eqs. (14.40) and (14.41).

The mathematical problem is then a problem of the calculus of variations with the differential Eq. (14.40) as an auxiliary condition. This can again be solved by using the method of the Euler-Lagrange multiplier $\lambda(t)$.[1] That is,

$$\int_0^{t_1} F(y,\dot{y},\ddot{y};z,\dot{z},\ddot{z};t)\, dt = \text{min.} \tag{14.42}$$

with

$$F = f(y) + \lambda(t) g(y,\dot{y},\ddot{y};z,\dot{z},\ddot{z};t) \tag{14.43}$$

The only novel feature of the problem is the introduction of the time-varying multiplier $\lambda(t)$. The problem of Eq. (14.42) is exactly the same as that of Eq. (14.30). Therefore all the equations developed in the last section can be used. However, we now have three unknowns: y, z, and λ. The three equations are Eqs. (14.34) and (14.40) with F defined by Eq. (14.32). Equation (14.40) is intrinsic to the physical system and is thus automatically satisfied by the controlled system. What has to be artificially enforced is Eq. (14.34). That system of two equations then forms the basis of the operation of the control computer. A properly constructed control computer then takes information about the essential output y, digests it, and then generates a continuous signal for z. This signal of z, when fed into the controlled system, then forces the system to behave according to the specification of Eq. (14.41).

[1] See for instance O. Bolza, "Vorlesungen über Variationsrechnung," Chap. 11, Teubner, 1909.

14.8 Comparison of Concepts of Control Design. In the last chapter and in this chapter, we have discussed methods of designing control systems when the performance is specified quite rigidly. The method of the last chapter, based upon the perturbation theory, is applicable to linear systems with time-varying coefficients. The method of this chapter is even more general in that the system to be controlled may itself be nonlinear. For such general systems, these newer methods are the only available tools for designing the control; and the resultant complicated control systems involving electromechanical computers seem to be the only logical solution. However, the methods of the last two chapters are equally applicable to the simpler physical systems treated in the earlier chapters, *i.e.*, linear systems with constant coefficients. For such simpler systems, then, we have two different general approaches to the control problem. It is illuminating to compare these two design concepts.

The engine-control problem was treated in Chap. 5, using the conventional principles of servomechanism. In this chapter, we have discussed almost the same problem with the new design method with specified performance criteria. One point is immediately clear: the control system designed by the older method is linear, and the control component can be a simple *RC* circuit; while the control system designed by the new method is nonlinear, and the essential control unit is a computer which even in its simplest form is much more complicated than an *RC* circuit. But this complication is not introduced without a gain: while the control system based upon the principle of the conventional servomechanism may be entirely satisfactory in performance, the control system involving a nonlinear computer is guaranteed to give the optimum performance—no other control system can be better under the same design specification. But this comparison has meaning only if we know specifically the desired performance. For instance, if we do not know exactly what the temperature integral in Eq. (14.18) is, we cannot apply the method discussed in this chapter at all. On the other hand, to design a satisfactory control system according to the older concepts of servomechanisms requires no such sharply defined specification.

Of course, strict insistence on the best performance must come after a clear understanding of what constitutes optimum control behavior. Therefore, when we do want an optimum control system, we naturally should have the information to define sharply the design criteria. From this point of view, then, the newer principles of the last chapters for control systems with specified performance certainly go one step beyond the conventional theory of servomechanisms and are principles for more advanced control design. That the more advanced control system should also be more complicated is to be expected.

CHAPTER 15

OPTIMALIZING CONTROL

In the previous chapters we have discussed the design principles of control systems with increasing degrees of generality and complexity. However, one basic assumption was made throughout the treatment: the properties and characteristics of the system to be controlled were always assumed to be known. In the case of conventional linear servomechanisms, the transfer functions of the servos and other components are specified before the design. In the case of linear systems with time-varying coefficients, we take the example of the guidance system for the long-range rocket vehicle. There, the dynamic and the aerodynamic properties of the rocket were determined previous to the design of the control system. In the case of the general system control according to specified criteria treated in the last chapter, the response of the system to variations in the controlled input is again predetermined. The control design is thus based upon this knowledge of the properties of the system. The feedback merely conveys the information on the state of the output to the "computer." The computer then uses its built-in knowledge of the system properties to generate the "intelligent" control signal.

In this chapter, we wish to relax even this seemingly elementary requirement for control design. We wish to introduce the principle of *continuously sensing and continuously measuring control systems* where no exact knowledge of the properties of the controlled system is necessary for the design. Instead, the properties of the controlled system are measured during the control process. In particular, we shall discuss a simple case of such control systems, the *optimalizing control.*

15.1 Basic Concept. No matter what degree of accuracy we can obtain from the control computer, the accuracy of the controlled behavior of a system is still dependent upon the accuracy of the information used in its design. If we determine the properties of the controlled system before we know the over-all design of the control system, as is tacitly assumed in the previous chapters, then extreme accuracy of the controlled behavior cannot be expected for two reasons: First, the manufacturing process always introduces small differences into supposedly identical objects. For instance, the wing of a rocket cannot be identical to the wing of the rocket model on which the wind-tunnel tests are made to

determine the aerodynamic properties. Therefore the aerodynamic properties of the rocket in reality must differ slightly from the test results. Secondly, any engineering system is subject to small variations with respect to time. This may be due to the normal deterioration of the system caused by wear and fatigue, or it may be due to the drift of conditions in the environment in which the system operates. In short, the properties of an engineering system can never be known *exactly* prior to the instant of actual operation of the system. Therefore, when great accuracy of the controlled behavior is a necessity, we must use the principle of a continuously sensing control system.

Nor is the required accuracy of control the only reason for changing our concept of control; very often the fact that large unpredictable variations of the system properties may occur forces us to use the continuously sensing system. We have already introduced this principle in connection with the disturbance effects of atmospheric changes in the guidance problem of the long-range rocket. There we used the dynamic behavior of the rocket itself as an instrument for continuous measurement of these effects. But a more illuminating example is the flight of an airplane through icing weather conditions. The deposition of ice on the surfaces of the wing and the fuselage changes the shape of these airplane components. Moreover, the exact manner of deposition is somewhat variable and cannot be predicted with accuracy. Therefore the aerodynamic properties of the airplane can be profoundly altered, and altered in an unpredictable way, by ice. Even worse, the changes all tend to degrade the performance of the airplane, *i.e.*, decrease the number of miles that can be flown with one gallon of gasoline. We are thus interested in knowing the combination of engine throttle, engine rpm, and airplane trim that will give the maximum miles per gallon of gasoline, because we should fly the airplane at the optimum condition to conserve the strained fuel load. But just at this critical situation, our prior knowledge of the airplane performance is rendered useless by the ice deposition. Hence the only solution to this problem of cruise control in adverse circumstances is an automatic sensing and measuring control system, *i.e.*, an optimalizing system which automatically holds the airplane at the measured optimum operating conditions.

Of course, a skilled human operator controls the performance of a machine on the optimalizing principle. He watches the instrument readings of the inputs and outputs of the machine, and then uses his knowledge and experience to decide in what directions the controls should be adjusted. The adjusted inputs bring new output readings, which have to be interpreted by the operator to determine whether the optimum operating condition is reached or exceeded. New adjustments of the controls will have to be made. The continuous adjustment of inputs is

the sensing process, and the reading of the outputs is the feedback. However, manually controlled optimalizing systems are necessarily slow in response, and, for complicated systems, human skill, no matter how well developed, is not sufficient. Automatic optimalizing control was conceived by C. S. Draper, Y. T. Li, and H. Laning, Jr.[1] Its application to cruise control of an airplane was discussed by J. R. Shull.[2]

15.2 Principles of Optimalizing Control. The heart of an optimalizing control system is the nonlinear component which characterizes the optimum operating conditions. For simplicity of discussion, we shall assume that this basic component has a single input and a single output. For the time being, we shall also neglect any time effects and assume that the output is determined only by the instantaneous value of input. Since there is an optimum operating point, the output as a function of input has a maximum at y_o and x_o, as shown in Fig. 15.1. It is convenient to refer the output and the input to the optimum point, and thus put the physical input as $x + x_o$ and the physical output as $y^* + y_o$. The optimum point is then the point $x = y^* = 0$. The purpose of an optimalizing control is then to search out this optimum point and to keep the system in the immediate neighborhood of this point. In this neighborhood, the relation between x and y^* can be represented as

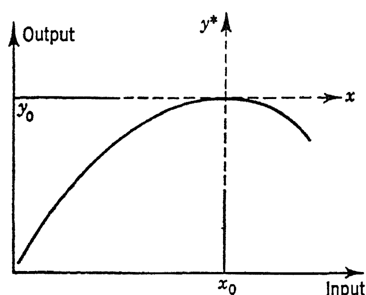


FIG. 15.1

$$y^* = -kx^2 \tag{15.1}$$

The simplest method of obtaining an optimalizing system, in concept, is as follows. Let us assume that we start with a negative input, *i.e.*, an input smaller than the optimum input, and that we increase this input at a constant time rate as shown in Fig. 15.2*a*. The corresponding output y^* will first increase, then reach the optimum value, and start to decrease, as shown in Fig. 15.2*b*. The time derivative of y^*, dy^*/dt, is then first positive, decreases to zero at the point 1 (Fig. 15.2*c*), and becomes negative after that point. At the point 2, the value of dy^*/dt reaches the critical magnitude designed into the control system so that the direction of variation of the input is reversed, and the input x now decreases with the same constant rate. y^* now increases again, and dy^*/dt jumps to

[1] Y. T. Li, *Instruments*, **25**, 72–77, 190–193, 228, 324–327, 350–352 (1952). C. S. Draper and Y. T. Li, "Principles of Optimalizing Control Systems and an Application to Internal Combustion Engine," ASME Publications (1951).

[2] J. R. Shull, *Trans IRE* (Electronic Computers), December, 1952, pp. 47–51.

positive values. At the point 3, the output reaches its maximum value, and dy^*/dt becomes zero again. At the point 4, dy^*/dt again reaches the critical value, and the drive for the input variation is again reversed. This process repeats itself, and the behavior of the system is periodic. The system is said to hunt around the optimum point. The period T^* is the *hunting period.* The minimum value of the output is Δ^* and is called the *hunting zone* of the output y^*. Because of the parabolic

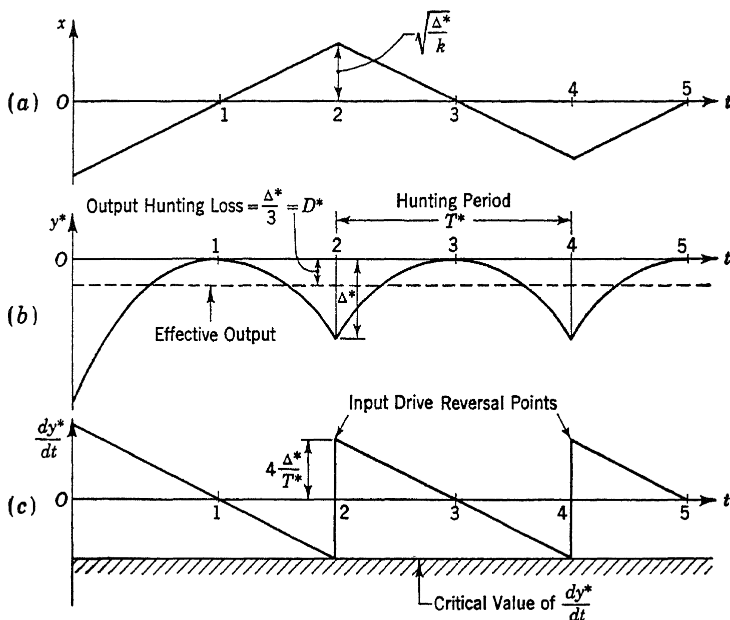


FIG. 15.2

relation of Eq. (15.1), the average value of the output with hunting is $\frac{1}{3}\Delta^*$ lower than the optimum output. This difference is a loss, the *hunting loss* D^*, and is the price to be paid for the control. Thus

$$D^* = \tfrac{1}{3}\Delta^* \tag{15.2}$$

Other characteristics of the system can be calculated in terms of Δ^* and T^*: By using Eq. (15.1), the extreme values of the input are $\pm\sqrt{\Delta^*/k}$. The rate of change of the input is thus $2\sqrt{\Delta^*/k}/T^*$. The critical value of the time rate of output change is $-4\Delta^*/T^*$. Therefore, if we fix the hunting zone Δ^* or the hunting loss D^*, and the hunting period T^*, the system is specified. The essential elements of such an optimalizing system are the test variations of the input, the output detection and differentiating device, and the switching of the input drive at the predetermined magnitudes. The sensing and searching for the

optimum point are accomplished by the forced input variations. But the constantly changing input also causes a small loss D of the output. It is desirable to make the hunting zone Δ^* small. But small Δ^* also reduces the magnitude of the critical dy^*/dt for input drive reversal. This then increases the danger of accidental input drive reversals caused by the unavoidable disturbances or noise in the system. It is apparent that if the system is displaced from the optimum point, the recovery time is directly proportional to the hunting period T^*. A short hunting period is thus desirable. But if T^* is made too small, it will be difficult to differentiate the output variation from the hunting operation and other random disturbances. We shall discuss this point again in the next section.

The testing input variation can also be a continuous function of time instead of the saw-tooth curve of Fig. 15.2a. For instance, we can make the input x be a combination of a slowly varying part x_a and a sinusoid of constant amplitude a and frequency ω. Thus

$$x = x_a + a \sin \omega t \tag{15.3}$$

Then, according to Eq. (15.1), the corresponding output y^* is

$$y^* = -k\left(x_a^2 + \frac{a^2}{2}\right) - 2kx_a a \sin \omega t + \frac{ka^2}{2} \cos (2\omega t) \tag{15.4}$$

This output signal can be fed to a band-pass filter to remove both the slowly varying first term and the double harmonic of the third term. The filtered signal is thus $-2kx_a a \sin \omega t$. Now this signal and the sinusoidal signal $a \sin \omega t$ are combined in a rectifying multiplier which multiplies these signals to give

$$-2kx_a a^2 \sin^2 \omega t = -kx_a a^2[1 - \cos (2\omega t)] \tag{15.5}$$

and then we again remove the double harmonic. We have then, finally, the signal $-ka^2x_a$. This signal can be used to vary the part x_a of the input, such that

$$\alpha \frac{dx_a}{dt} = -ka^2 x_a \tag{15.6}$$

Then x_a tends to zero with a decay time $2T^*$,

$$2T^* = \frac{\alpha}{ka^2} \tag{15.7}$$

Because of the parabolic relation between input and output, the decay time constant for the output is T^*. Therefore such a control system will also search out the optimum point and approach it asymptotically. The operation of this optimalizing control with continuous test signal is

shown in Fig. 15.3. Fig. 15.3*c* shows the filtered output signal, and Fig. 15.3*d* indicates the effects of the rectifying multiplier.

When the system is operating near the optimum point, because of the sinusoidal oscillation of the input, the output is $-ka^2 \sin^2 \omega t$. Therefore

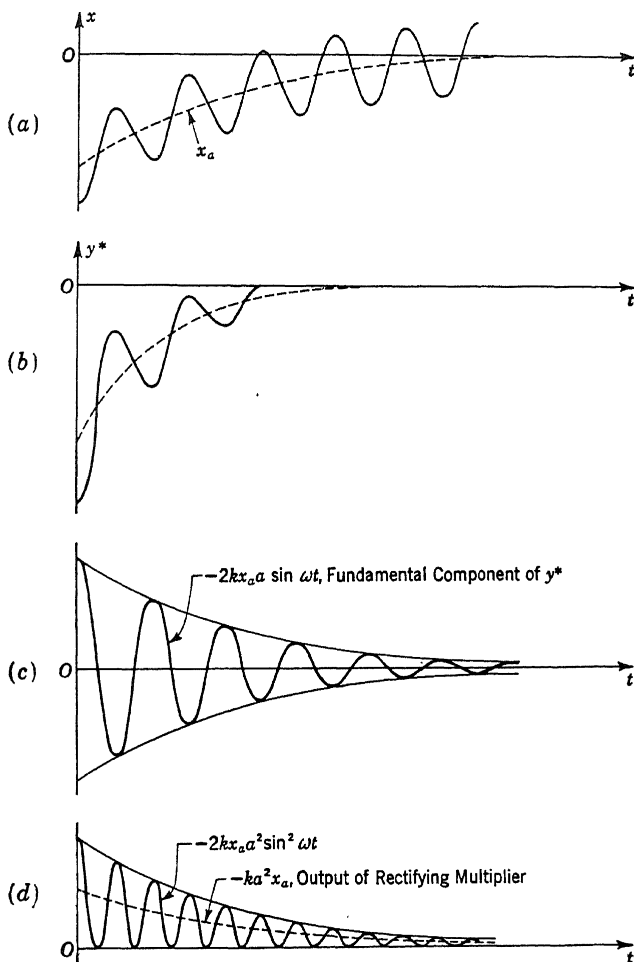


FIG. 15.3

there is again an output loss $D^* = ka^2/2$. For low losses, the amplitude of the test input should be small. But this is again limited by considerations on incidental disturbances and noise of the system. The hunting zone Δ^* of the output is ka^2, and thus

$$D^* = \tfrac{1}{2}\Delta^* \tag{15.8}$$

Equation (15.7) shows that the design constant α of the input drive is determined by the time constant T^* and D^* according to the relation

$$\alpha = 4D^*T^* = 2\Delta^*T^* \tag{15.9}$$

The rectified part of the signal given by Eq. (15.5), $-ka^2x_a$, obtained from the test input of Eq. (15.3), is actually a measure of the deviation of the input from the optimum input. The continuous drive of the input according to Eq. (15.6) is but one of the many possible uses of this signal. Evidently, this signal can also be used to give a saw-tooth variation of x_a by using a constant rate variation of input with a superimposed sinusoidal oscillation, and by reversing the input drive at critical values of the signal $|-ka^2x_a|$. The hunting operation of this optimalizing control then consists of two separate frequencies, a low-frequency component in the variation of x_a, and a high-frequency component produced by the sinusoidal input oscillation.

15.3 Considerations on Interference Effects. The previous discussions on the idealized optimalizing controls have shown the importance of reducing the amplitude of the variation of test input and the time

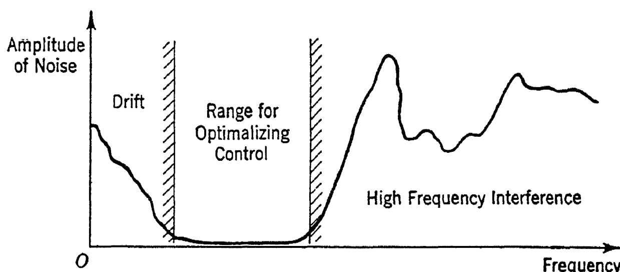


Fig. 15.4

constant T^*. However the actual design is limited in this respect by the ubiquitous noise and interference in the physical system. In order to measure effectively the output variation as a result of input variation for testing the distance to the optimum point, it is necessary for this output variation with time to be made up of frequency components that may with certainty be distinguished from output variations due to noise and interference effects. The relative amplitudes of the interference output frequency components can be plotted as a function of the frequency. This output interference spectrum generally has a low-frequency part, the drift interferences, and a high-frequency part, as shown in Fig. 15.4. Between these two parts, there is generally a range of frequency relatively free of the noise effects. If the optimalizing control is designed with test variations with frequencies in this range, the amplitude of these test variations can be made small without fear of losing them in the background of interferences. In general then, the test func-

tion must be made up of input variations that are fast enough to be distinguishable from drift interference and at the same time slow enough to prevent confusion with high-frequency noise.

These considerations on the noise effects point out the difficulties in both types of optimalizing controls discussed in the previous section. The first type, with saw-tooth input drive, uses the time derivative of the output as the control signal. If there is random interference in the output, then the relative amplitudes of the high-frequency components will be increased by taking the time derivative; thus the range of frequency available for optimalizing control will be reduced. This is a serious disadvantage. The second type of optimalizing control, using the continuous sinusoidal test function, requires a wide band of noise-free frequencies, because in addition to the variation of the level of input, x_a, there is the sinusoidal variation with its own higher frequency ω. Thus if the system to be controlled has only a narrow range of noise-free frequencies, the two optimalizing controls so far discussed are not easily applicable. A better system is the so-called *peak-holding optimalizing control* to be discussed in the next section.

15.4 Peak-holding Optimalizing Control. The input variation for a peak-holding optimalizing control is the same as for the first type of optimalizing control studied here, a constant rate variation with periodic reversals. The essential improvement here lies in the method of generating the drive-reversal signal. Reversal of the input drive should occur when the output has passed its maximum and has decreased to a value approaching the hunting zone limit. This fact itself is used as the condition for producing the input drive reversal for the peak-holding optimalizing control. Essentially this could be accomplished as follows. The output y^* is measured as a voltage. This voltage is applied to a condenser through a gate allowing only charging but not discharging. Then the voltage of the condenser follows the output y^* until the maximum value is reached. When the input x is increased beyond the optimum value, the output y^* decreases. But the condenser voltage will remain at its maximum value, and a voltage difference v exists between the condenser and the indicating voltage for the output. This voltage difference v builds up to the hunting zone limit Δ^*. At this point a relay is triggered to reverse the input drive, and at the same time the condenser is discharged to a voltage equal to the instantaneous y^*. Thus the operation of this optimalizing control can be represented by Fig. 15.5.

The relation between the hunting zone Δ^* and the hunting loss D^* is the same as given by Eq. (15.2). The extreme values of the input are still $\pm\sqrt{\Delta^*/k}$, and the input rate is again $2\sqrt{\Delta^*/k}/T^*$. It is seen that the peak-holding optimalizing control has only one essential

frequency of output, determined by the hunting period T^*, and no differentiation of the output is used. It is thus particularly adapted to a system with a narrow noise-free frequency range. In fact a further improvement in this direction could be made by basing the input drive reversal not directly on v, the voltage difference between condenser and output indicator, but the integral of v with respect to time. Then

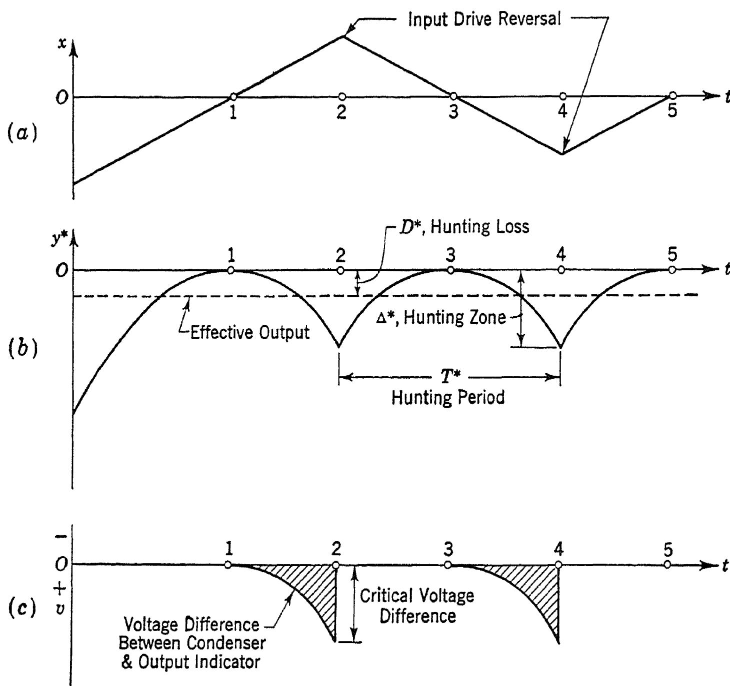


FIG. 15.5

the high-frequency interference effects will be suppressed, and the hunting zone and hunting loss can be further reduced without accidental input drive reversal.

15.5 Dynamic Effects. In the discussions of previous sections, we have assumed that the input-output relation is specified by Eq. (15.1) and is independent of the rate of change of the input or the higher time derivatives of input. This is true if the response of the output to input variation is instantaneous without the slightest time delay. In any physical system, this is not possible; there are always the inertial and other dynamic effects. We have then to consider the output y^* given by Eq. (15.1) as the fictitious "potential output" but not the actual output y measured by the output-indicating instrument. y^* is equal to y only when the time constant T^* of the optimalizing control approaches

infinity. The relation between y^* and y is determined by the dynamic effects. But we have seen previously that such dynamic effects can be closely approximated by a linear system. If the optimalizing control is to be applied to an internal-combustion engine, as was done by C. S. Draper and Y. T. Li, the potential output is essentially the indicated mean effective pressure of the engine, while the actual output is the brake mean effective pressure of the engine. The dynamic effects here are mainly due to the inertia of the piston, the crankshaft, and other moving parts of the engine. For small changes in the operating conditions of the engine, such dynamic effects can be represented as a linear differential equation with constant coefficients. Since we have set the reference level of input and output at the optimum input x_o and the optimum output y_o, the physical potential output is $y^* + y_o$, and the physical indicated output is $y + y_o$. Thus the relation between the physical potential output and the physical indicated output can be written as an operator equation

$$(y + y_o) = F_o\left(\frac{d}{dt}\right)(y^* + y_o)$$

where F_o is generally a quotient of two polynomials in the operator d/dt. In the language of Laplace transforms, $F_o(s)$ is the transfer function. Let us call the linear system which transforms the potential output to the indicated output used for controlling the input variation, the *output linear group* of the optimalizing control. Then $F_o(s)$ is the transfer function of the output linear group. By implication, however, when the dynamic effects are negligible, or when $s = 0$, the potential output is equal to the indicated output. Therefore we have the condition that

$$F_o(0) = 1 \tag{15.10}$$

Then the operator equation between the potential output and the indicated output can be simplified because y_o is a constant. That is,

$$y = F_o\left(\frac{d}{dt}\right)y^* \tag{15.11}$$

In a similar manner, we can introduce a "potential input" x^* which is actually the forcing function generated by the optimalizing control system but not the actual input x. The relation between x and x^* is determined by the inertial and dynamic effects of the input drive system. This input drive system we shall call the *input linear group* of the optimalizing control. And the operator equation between the potential input x^* and the actual input x is

$$x = F_i\left(\frac{d}{dt}\right)x^* \tag{15.12}$$

$F_i(s)$ is thus the transfer function of the input linear group. Similarly to Eq. (15.10), we have

$$F_i(0) = 1 \tag{15.13}$$

Thus the block diagram of the complete optimalizing control system can be drawn as shown in Fig. 15.6. The nonlinear components of the system are the optimalizing input drive and the controlled system itself.

The general relation between the input x and the output y is then determined by the system of Eqs. (15.1), (15.11), and (15.12) and by the particular optimalizing input drive adopted. For instance, if the optimalizing input drive is of the peak-holding type discussed in the last

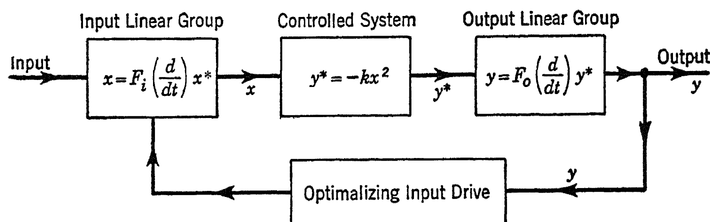


FIG. 15.6

section, then the potential input x^* is a saw-tooth curve with period $2T$ and amplitude a, as shown in Fig. 15.7a. Let

$$\omega_0 = \frac{2\pi}{T} \tag{15.14}$$

Then x^* can be expanded into a Fourier series,

$$\begin{aligned} x^* &= \frac{8a}{\pi^2} \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n+1)^2} \sin\left[(2n+1)\frac{\omega_0 t}{2}\right] \\ &= \frac{8a}{\pi^2} \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n+1)^2} \frac{1}{2i} \left[e^{\frac{2n+1}{2}i\omega_0 t} - e^{-\frac{2n+1}{2}i\omega_0 t}\right] \end{aligned} \tag{15.15}$$

According to the general relation of Eq. (2.16), the actual input x, given by Eq. (15.12), can then be calculated as

$$\begin{aligned} x = \frac{8a}{\pi^2} \sum_{n=0}^{\infty} \frac{(-1)^n}{2i(2n+1)^2} \bigg[& F_i\left(\frac{2n+1}{2} i\omega_0\right) e^{\frac{2n+1}{2}i\omega_0 t} \\ & - F_i\left(-\frac{2n+1}{2} i\omega_0\right) e^{-\frac{2n+1}{2}i\omega_0 t}\bigg] \end{aligned} \tag{15.16}$$

The potential output y^* is given by Eq. (15.1). Using Eq. (15.16), we have

$$y^* = \frac{16a^2k}{\pi^4} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{(-1)^{n+m}}{(2n+1)^2(2m+1)^2}$$
$$\left[F_i\left(\frac{2n+1}{2} i\omega_0\right) F_i\left(\frac{2m+1}{2} i\omega_0\right) e^{(n+m+1)i\omega_0 t}\right.$$
$$- F_i\left(\frac{2n+1}{2} i\omega_0\right) F_i\left(-\frac{2m+1}{2} i\omega_0\right) e^{(n-m)i\omega_0 t}$$
$$- F_i\left(-\frac{2n+1}{2} i\omega_0\right) F_i\left(\frac{2m+1}{2} i\omega_0\right) e^{-(n-m)i\omega_0 t}$$
$$\left.+ F_i\left(-\frac{2n+1}{2} i\omega_0\right) F_i\left(-\frac{2m+1}{2} i\omega_0\right) e^{-(n+m+1)i\omega_0 t}\right] \qquad (15.17)$$

By again applying Eq. (2.16), we have, finally, the indicated output y, according to Eq. (15.11),

$$y = \frac{16a^2k}{\pi^4} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{(-1)^{n+m}}{(2n+1)^2(2m+1)^2}$$
$$\left\{F_o[(n+m+1)i\omega_0]F_i\left(\frac{2n+1}{2} i\omega_0\right) F_i\left(\frac{2m+1}{2} i\omega_0\right) e^{(n+m+1)i\omega_0 t}\right.$$
$$- F_o[(n-m)i\omega_0]F_i\left(\frac{2n+1}{2} i\omega_0\right) F_i\left(-\frac{2m+1}{2} i\omega_0\right) e^{(n-m)i\omega_0 t}$$
$$- F_o[-(n-m)i\omega_0]F_i\left(-\frac{2n+1}{2} i\omega_0\right) F_i\left(\frac{2m+1}{2} i\omega_0\right) e^{-(n-m)i\omega_0 t}$$
$$+ F_o[-(n+m+1)i\omega_0]F_i\left(-\frac{2n+1}{2} i\omega_0\right) F_i$$
$$\left.\left(-\frac{2m+1}{2} i\omega_0\right) e^{-(n+m+1)i\omega_0 t}\right\} \qquad (15.18)$$

Equations (15.17) and (15.18) clearly indicate that the hunting period T of the output is only half of the period of the input variation. This is, of course, to be expected from the basic parabolic relation of input to output.

The average value of y with respect to time, being here referred to the optimum output y_o, gives the hunting loss D. Equation (15.18) shows that this average value is the sum of second and third terms of that equation with $n = m$. Thus, noting Eq. (15.10),

$$D = \frac{32a^2k}{\pi^4} \sum_{n=0}^{\infty} \frac{1}{(2n+1)^4} \left| F_i\left(\frac{2n+1}{2} i\omega_0\right)\right|^2 \qquad (15.19)$$

This equation can be easily checked by observing that when the dynamic effects are absent, $F_i \equiv 1$; then the series can be easily summed, and

$D = D^* = a^2k/3 = \Delta^*/3$, as required by Eq. (15.2). Equation (15.19) also shows that the average output and the hunting loss are independent of the output linear group. This is, of course, to be expected, since the level of output is determined by the input x, and is not influenced by the dynamic effects of the output linear system. Only detailed time variation of the output is modified by the dynamics of the output linear group. In the case of an internal-combustion engine, the output level is the

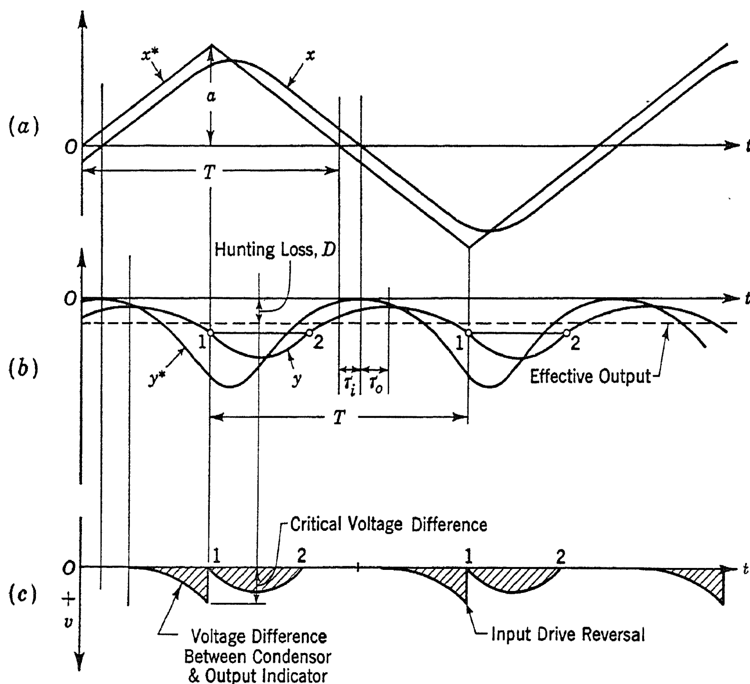


FIG. 15.7

power of the engine. The dynamics of the output linear group is determined by the inertia of the moving parts. The power of the engine is certainly independent of the inertia of the moving parts.

The numerical calculation of the output y from Eq. (15.18) for general input and output transfer functions is rather difficult. However, any practical design of an optimalizing control usually has a rather long hunting period T to avoid the high-frequency interference. Then the dynamic effects, although not negligible, are not large. In other words, we can assume that the ratios of the time constants of the input and output linear groups to the hunting period are small, and carry out the analysis accordingly. For instance, if the input linear group is approximated by a first-order system with the time constant τ_i, *i.e.*,

$$F_i(s) = \frac{1}{1 + \tau_i s} \tag{15.20}$$

and if τ_i is small in comparison with T, then the nondimensional quantity $\tau_i\omega_0$ is also small. Under this condition, the first few corresponding harmonics of the series of Eqs. (15.15) and (15.16) will have practically the same amplitudes, according to Eq. (3.14). The only difference between these corresponding lower harmonics in x^* and in x is a time shift of magnitude τ_i. Therefore, for regions of x^* and x away from the drive reversal points where the curvatures of $x^*(t)$ and $x(t)$ curves are small and where the values are determined mainly by the first few harmonics, the $x(t)$ curve merely lags the $x^*(t)$ curve by τ_i without change in magnitude. In going from x^* to x, the sharp corners will be rounded off, but the general shape of the curves remains unchanged, as shown in Fig. 15.7*a*. If the output linear group is also approximated by a first-order system with a characteristic time τ_o, then similar considerations will show that in going from y^* to y, the pattern of the curves remains the same, but y lags behind y^* by τ_o. This fact is shown in Fig. 15.7*b*.

With the input transfer function given by Eq. (15.20), the hunting loss can be calculated by Eq. (15.19). Thus

$$\begin{aligned} D &= \frac{32a^2k}{\pi^4} \sum_{n=0}^{\infty} \frac{1}{(2n+1)^4} \frac{1}{1 + (2n+1)^2(\tau_i\omega_0/2)^2} \\ &= \frac{32a^2k}{\pi^4} \left[\sum_{n=0}^{\infty} \frac{1}{(2n+1)^4} - \left(\frac{\tau_i\omega_0}{2}\right)^2 \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} \right. \\ &\qquad \left. + \left(\frac{\tau_i\omega_0}{2}\right)^4 \sum_{n=0}^{\infty} \frac{1}{1 + (2n+1)^2(\tau_i\omega_0/2)^2} \right] \end{aligned}$$

But we have

$$\sum_{n=0}^{\infty} \frac{1}{(2n+1)^4} = \frac{\pi^4}{96} \qquad \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} = \frac{\pi^2}{8}$$

and, by using the well-known expansion for the hyperbolic cotangent cited on p. 90,

$$\sum_{n=0}^{\infty} \frac{1}{1 + (2n+1)^2(\tau_i\omega_0/2)^2} = \frac{\pi}{\tau_i\omega_0} \left[\coth \frac{2\pi}{\tau_i\omega_0} - \frac{1}{2} \coth \frac{\pi}{\tau_i\omega_0} \right]$$

Therefore, by substituting the period T from Eq. (15.14), we have, finally,

$$D = \frac{a^2k}{3}\left[1 - 12\left(\frac{\tau_i}{T}\right)^2 + 48\left(\frac{\tau_i}{T}\right)^3\left(\coth\frac{T}{\tau_i} - \tfrac{1}{2}\coth\frac{T}{2\tau_i}\right)\right] \quad (15.21)$$

When the time lag τ_i is much smaller than T, the hyperbolic cotangents are approximately equal to unity. Then

$$D \approx \frac{a^2k}{3}\left[1 - 12\left(\frac{\tau_i}{T}\right)^2 + 24\left(\frac{\tau_i}{T}\right)^3\right] \qquad \frac{\tau_i}{T} \ll 1 \quad (15.22)$$

Since the input amplitude a can be expressed in terms of the input drive speed and the period T, Eqs. (15.21) and (15.22) give the hunting loss D in terms of input drive speed and the hunting period T, for a peak-holding optimalizing control with a first-order input linear group of lag τ_i. These equations apparently indicate that a decrease of hunting loss is caused by the lag of the input linear group. However, it is deceptive: for a given critical voltage difference v for input drive reversal, determined by considerations of noise and interference, the hunting period T and hence the amplitude a will be much larger with the time lags τ_i and τ_o than without the time lags. The net result is an increase instead of a decrease in hunting loss.

15.6 Design for Stable Operation. Stability of any control system means that the design performance of the system will be obtained even with the presence of internal and external disturbances. We have seen how this requirement is satisfied in the case of conventional servomechanisms and other more general control systems in the preceding chapters. For optimalizing control systems, the essential part of the operation is the proper coordination of the input drive with the output behavior, so that the output stays within a close neighborhood of the optimum. This operation must not be influenced by internal and external disturbances. When this is achieved by a good design of the system, we have stable operation.

For a peak-holding optimalizing control, we have described the input drive signal as the result of the charging and discharging of a condenser by a voltage representing the magnitude of the output. The input drive-reversal signal is given when the voltage difference v between the output indicating voltage and the condenser builds up to a critical value because of decreasing output. At the instant of reversal of the input drive, the condenser voltage is reset by discharging to the output indicating voltage at the same instant. When there are dynamic effects, the time lags of the input and output linear groups will cause the output to continue to decrease even after the signal for input drive reversal has been given. Then the voltage v again builds up and will be removed only after the output has risen to a value equal to that at the instant of signaling for input drive reversal. This is shown in Fig. 15.7c. This

positive spurious voltage v between the instants 1 and 2 (Fig. 15.7) is undesirable because of the danger of tripping the input drive control during this wrong interval of time. To greatly reduce the positive spurious voltage, the condenser voltage is reset at the instant of input

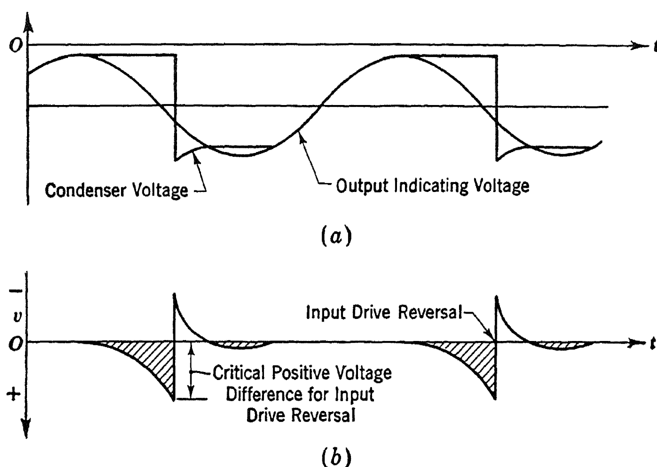


FIG. 15.8

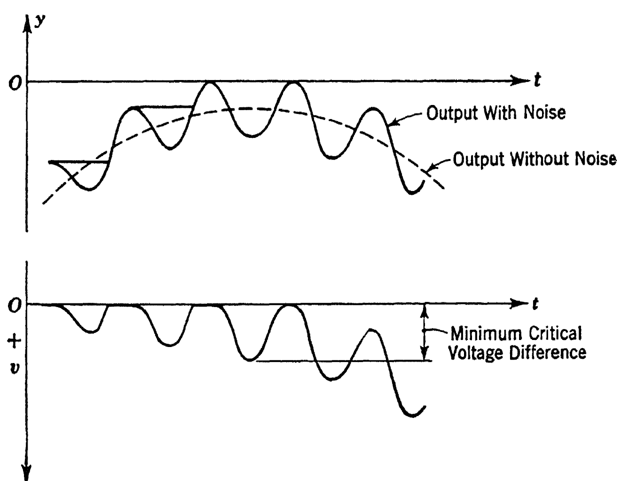


FIG. 15.9

drive reversal to a lower voltage than the instantaneous output indicating voltage. Thus during an interval of time after the reset, the condenser is charged by the output indicating voltage. The capacity of the condenser and the resistance of the electric circuit are so chosen as to make the voltage of the condenser approach that of the output

indicating voltage when the output is increasing again. The variation of voltages is shown in Fig. 15.8. The dangerous positive spurious voltage difference is thus greatly reduced (Fig. 15.8*b*), and the stability of the control system improved.

We have already stated that the reduction of the hunting zone and hunting loss is limited by the interference and noise of the system. This is again a problem of stable operation: we do not wish to have false signals for reversing the input drive caused by interference. Such false signals will occur if the critical voltage difference for input drive reversal is too small. This can be shown by Fig. 15.9 where the output y contains a high-frequency sinusoidal noise. It is easily seen that if the critical voltage difference is too small, the noise will trip the input drive in an erratic manner. For stable operation, the critical voltage difference must be larger than the amplitude of the interference. Thus the hunting loss of an optimalizing control cannot be less than the interference or noise of the system. Of course, if the interference actually were a pure high-frequency sine wave of constant amplitude, as shown in the figure, it could be removed by a filter, and a much smaller hunting zone could be used. In fact, if the interference or noise has any definite pattern at all, we can design a proper filter to ameliorate this limitation.

CHAPTER 16

FILTERING OF NOISE

In all the previous discussions with the exception of the last chapter, we have tacitly assumed that the control system does not generate noise and interference, so that theoretically there is no limit to the accuracy of control. In the last chapter, we have shown that noise and interference in fact obscure the output signal used for the optimalizing control and are the fundamental design restrictions in such control systems. But noise and interference are present in any engineering system, because even the "perfect" systems have thermodynamic fluctuations. Their effects on the control system are negligible only if the signal is relatively strong in comparison to the interference. For the optimalizing system, to minimize hunting loss of the output, we design for "weak" signals, and thus the problem of noise is of paramount importance. In general then, whenever the control signal is weak in comparison to interference, the effects of noise and interference cannot be neglected.

The disturbing influence of noise on the control system can be minimized by introducing a proper device which will "filter" out the noise as much as possible without reducing the strength of the signal. This subject of noise filtering is the theme of this chapter. We shall first give a discussion of the theory of optimum linear filters developed by N. Wiener[1] and A. Kolmogoroff.[2] Later parts of this chapter will treat the various applications and extensions of this very powerful theory. The concepts and mathematical tools introduced in Chap. 9 on random inputs are very useful in our present discussion.

16.1 Mean-square Error. Let $f(t)$ be the control signal and $n(t)$ be the noise. The input $x(t)$ to the filter is thus

$$x(t) = f(t) + n(t) \tag{16.1}$$

The output from the filter is $y(t)$, as shown in Fig. 16.1. If the filter is a linear filter and the differential equation for the output-input relation is a linear differential equation of constant coefficients, then the filter properties are completely determined by its transfer function $F(s)$.

[1] N. Wiener, "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications," John Wiley & Sons, Inc., New York, 1949.

[2] A. Kolmogoroff, *Bull. acad. sci. U.R.S.S., Ser. Math.*, **5**, 3–14 (1941).

When $F(s)$ is known, the response $h(t)$ of the filter to a unit impulse is given by Eq. (2.18). If $F(s)$ has poles only in the left-half s plane, we can write

$$h(t) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{st}F(s)\,ds = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t}F(i\omega)\,d\omega \tag{16.2}$$

The output $y(t)$ due to the input $x(t)$ of Eq. (16.1) is then

$$y(t) = \int_{-\infty}^{t} x(\eta)h(t - \eta)\,d\eta$$

assuming that the input extends far into the past. Let $t - \eta = \tau$; then

$$y(t) = \int_{0}^{\infty} x(t - \tau)h(\tau)\,d\tau \tag{16.3}$$

Let the desired output be $z(t)$, which is determined by the signal $f(t)$ and the desired impulse response $h_1(t)$, *i.e.*,

$$z(t) = \int_{0}^{\infty} f(t - \tau)h_1(\tau)\,d\tau \tag{16.4}$$

$h_1(t)$ can be calculated from the desired transfer function $F_1(s)$, and we

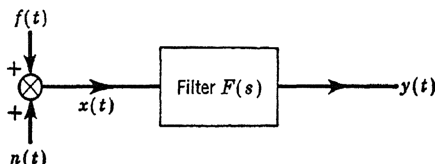


FIG. 16.1

have, similar to Eq. (16.2),

$$h_1(t) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{st}F_1(s)\,ds = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t}F_1(i\omega)\,d\omega \tag{16.5}$$

Since the actual output is not $z(t)$ but $y(t)$, the error $e(t)$ is their difference, and, according to Eqs. (16.3) and (16.4),

$$\begin{aligned} e(t) &= y(t) - z(t) \\ &= \int_{0}^{\infty} \{[f(t - \tau) + n(t - \tau)]h(\tau) - f(t - \tau)h_1(\tau)\}\,d\tau \end{aligned} \tag{16.6}$$

The square of the error is then

$$\begin{aligned} e^2(t) = \int_{0}^{\infty}\int_{0}^{\infty} &\{[f(t - \tau) + n(t - \tau)]h(\tau) - f(t - \tau)h_1(\tau)\} \\ &\{[f(t - \tau') + n(t - \tau')]h(\tau') - f(t - \tau')h_1(\tau')\}\,d\tau\,d\tau' \end{aligned} \tag{16.7}$$

We shall now make a very important assumption. Since the noise $n(t)$ is a random function, only its statistical properties can be specified.

In addition, we do not really know in advance what the signal $f(t)$ will be, but only its broad character. Therefore, even for the signal we can specify only the statistical properties. Then we can specify the statistical error to be measured by the assembly average of $e^2(t)$. In general this average is a function of the time instant t. But if we make the assumption that the random functions $f(t)$ and $n(t)$ are stationary as defined in Sec. 9.1, then $\overline{e^2}$ is independent of time. Furthermore, we shall assume that both the signal $f(t)$ and the noise $n(t)$ have zero mean value. Then it is evident from Eq. (16.6) that the mean value of the error $e(t)$ also vanishes. Now we can introduce the following correlation functions between the signal and the noise, entirely similar to the correlation functions of Sec. 9.2:

$$\left.\begin{aligned} \overline{f(t-\tau)f(t-\tau')} &= R_{ff}(\tau-\tau') \\ \overline{f(t-\tau)n(t-\tau')} &= R_{fn}(\tau-\tau') \\ \overline{n(t-\tau)f(t-\tau')} &= R_{nf}(\tau-\tau') \\ \overline{n(t-\tau)n(t-\tau')} &= R_{nn}(\tau-\tau') \end{aligned}\right\} \tag{16.8}$$

Here we take the general case of nonvanishing correlation between the signal and noise. Very often these *cross-correlation functions* R_{fn} and R_{nf} are zero, and only the *auto-correlation functions* R_{ff} and R_{nn} remain. The auto-correlation functions are symmetrical functions of the argument. The cross-correlation functions are not symmetrical functions, but they have the following relations according to their definitions

$$\left.\begin{aligned} R_{fn}(\tau'-\tau) &= R_{nf}(\tau-\tau') \\ R_{nf}(\tau'-\tau) &= R_{fn}(\tau-\tau') \end{aligned}\right\} \tag{16.9}$$

By using these definitions, the mean-square error $\overline{e^2}$ from Eq. (16.7) can be written as

$$\begin{aligned} \overline{e^2} = \int_0^\infty \int_0^\infty \{ & R_{ff}(\tau-\tau')[h(\tau)-h_1(\tau)][h(\tau')-h_1(\tau')] \\ & + R_{fn}(\tau-\tau')[h(\tau)-h_1(\tau)]h(\tau') + R_{nf}(\tau-\tau')h(\tau)[h(\tau')-h_1(\tau')] \\ & + R_{nn}(\tau-\tau')h(\tau)h(\tau')\}\, d\tau\, d\tau' \end{aligned} \tag{16.10}$$

This equation allows the calculation of the mean-square error from the correlation functions and the response to a unit impulse.

Control designers, however, prefer to make the analysis directly with the transfer functions $F(s)$ and $F_1(s)$. To do this, we introduce the Fourier transforms of the correlation functions. Let these Fourier transforms be $\Phi_{ff}(\omega)$, $\Phi_{fn}(\omega)$, $\Phi_{nf}(\omega)$, and $\Phi_{nn}(\omega)$, respectively, defined by the following equations:

$$\left.\begin{aligned}
\Phi_{ff}(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} R_{ff}(\tau)e^{-i\omega\tau}\,d\tau \\
\Phi_{fn}(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} R_{fn}(\tau)e^{-i\omega\tau}\,d\tau \\
\Phi_{nf}(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} R_{nf}(\tau)e^{-i\omega\tau}\,d\tau \\
\Phi_{nn}(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} R_{nn}(\tau)e^{-i\omega\tau}\,d\tau
\end{aligned}\right\} \tag{16.11}$$

Because of the fact that $R_{ff}(\tau)$ is a symmetrical function of τ, we can write

$$\Phi_{ff}(\omega) = \frac{1}{\pi}\int_0^{\infty} R_{ff}(\tau)(e^{i\omega\tau} + e^{-i\omega\tau})\,d\tau = \frac{2}{\pi}\int_0^{\infty} R_{ff}(\tau)\cos\omega\tau\,d\tau$$

By comparing this equation with Eq. (9.23), we see immediately that the function $\Phi_{ff}(\omega)$ is actually the power spectrum of the signal $f(t)$. Similarly, $\Phi_{nn}(\omega)$ is the power spectrum of the noise $n(t)$. Furthermore, since the cross-correlation functions are related as shown in Eq. (16.9), it is easily seen that the Fourier transforms are connected in a similar manner:

$$\left.\begin{aligned}
\Phi_{fn}(-\omega) &= \Phi_{nf}(\omega) \\
\Phi_{nf}(-\omega) &= \Phi_{fn}(\omega)
\end{aligned}\right\} \tag{16.12}$$

According to the Fourier integral theorem,[1] the inverse of Eq. (16.11) is

$$\left.\begin{aligned}
R_{ff}(\tau - \tau') &= \tfrac{1}{2}\int_{-\infty}^{\infty} \Phi_{ff}(\omega)e^{i\omega(\tau-\tau')}\,d\omega \\
R_{fn}(\tau - \tau') &= \tfrac{1}{2}\int_{-\infty}^{\infty} \Phi_{fn}(\omega)e^{i\omega(\tau-\tau')}\,d\omega \\
R_{nf}(\tau - \tau') &= \tfrac{1}{2}\int_{-\infty}^{\infty} \Phi_{nf}(\omega)e^{i\omega(\tau-\tau')}\,d\omega \\
R_{nn}(\tau - \tau') &= \tfrac{1}{2}\int_{-\infty}^{\infty} \Phi_{nn}(\omega)e^{i\omega(\tau-\tau')}\,d\omega
\end{aligned}\right\} \tag{16.13}$$

By substituting Eq. (16.13) into Eq. (16.10), we can obtain an equation for the mean-square error in terms of the transfer functions $F(s)$ and $F_1(s)$. For instance, the first part of Eq. (16.10) becomes

$$\begin{aligned}
&\tfrac{1}{2}\int_0^{\infty} d\tau \int_0^{\infty} d\tau' \int_{-\infty}^{\infty} d\omega\, \Phi_{ff}(\omega)e^{i\omega(\tau-\tau')}[h(\tau) - h_1(\tau)][h(\tau') - h_1(\tau')] \\
&\quad = \tfrac{1}{2}\int_{-\infty}^{\infty} d\omega\, \Phi_{ff}(\omega) \int_0^{\infty} [h(\tau) - h_1(\tau)]e^{i\omega\tau}\,d\tau \int_0^{\infty} [h(\tau') - h_1(\tau')]e^{-i\omega\tau'}\,d\tau'
\end{aligned}$$

But $F(s)$ and $F_1(s)$ are the Laplace transforms of $h(t)$ and $h_1(t)$, *i.e.*,

[1] See for instance Whittaker and Watson, "Modern Analysis," Sec. 6.31, p. 119, Cambridge-Macmillan, 1943.

$$\left.\begin{aligned} F(i\omega) &= \int_0^\infty h(t)e^{-i\omega t}\,dt \\ F_1(i\omega) &= \int_0^\infty h_1(t)e^{-i\omega t}\,dt \end{aligned}\right\} \tag{16.14}$$

Hence the first part of Eq. (16.10) can be written as

$$\frac{1}{2}\int_{-\infty}^{\infty} \Phi_{ff}(\omega)[F(i\omega) - F_1(i\omega)][F(-i\omega) - F_1(-i\omega)]\,d\omega$$

Other terms in Eq. (16.10) can be converted in a similar manner. We then obtain, finally,

$$\begin{aligned} \overline{e^2} = \frac{1}{2}\int_{-\infty}^{\infty} \{&\Phi_{ff}(\omega)[F(i\omega) - F_1(i\omega)][F(-i\omega) - F_1(-i\omega)] \\ &+ \Phi_{fn}(\omega)[F(-i\omega) - F_1(-i\omega)]F(i\omega) + \Phi_{nf}(\omega)F(-i\omega)[F(i\omega) - F_1(i\omega)] \\ &+ \Phi_{nn}(\omega)F(i\omega)F(-i\omega)\}\,d\omega \end{aligned} \tag{16.15}$$

The integrand within the brace can be considered the power spectrum of the error $e(t)$. The last term of the integrand is in fact the power spectrum of the filtered noise, according to Eq. (9.71). Evidently the first and the last terms are real. The second and the third terms are complex. However, because of Eq. (16.12), these two terms are complex conjugates, and therefore their sum is real.

16.2 Phillips's Optimum Filter Design. With the statistical properties of the signal and the noise given as the various correlation functions, the Fourier transforms can be computed by using Eq. (16.11). Then if the transfer function $F_1(s)$ is specified, the only function in the mean-square error integral of Eq. (16.15) yet to be fixed is the transfer function $F(s)$ of the filter. The optimum filter is then the filter which has a transfer function $F(s)$ such that the mean-square error is the minimum for specified Φ's and $F_1(s)$. A straightforward method for solving this problem of optimum filter design is to assume a reasonable form for $F(s)$, but with undetermined constants. Then $\overline{e^2}$ can be determined as a function of these undetermined constants by substituting the assumed $F(s)$ into Eq. (16.15). The constants are finally determined by requiring that the mean-square error must be a minimum with respect to these constants. The optimum filter design is thus reduced to a problem of finding the maximum or minimum of a known function. R. S. Phillips,[1] in fact, worked out such a theory of the optimum filter by taking $F(s)$ to be a ratio of two polynomials in s. This form of $F(s)$ is indeed a natural one, because our experience of linear systems indicates just such a choice. The actual calculation involved, however, is considerable. For this reason, the more elegant theory of Wiener and Kolmogoroff is

[1] R. S. Phillips, "Theory of Servomechanisms," MIT Radiation Laboratory Series, Vol. 25, Chap. 7, McGraw-Hill Book Company, Inc., New York, 1947.

generally preferred, and we shall not pursue the Phillips theory any further here.

16.3 Wiener-Kolmogoroff Theory. The theory of the optimum filter by Wiener and Kolmogoroff is based upon an application of the calculus of variations to the integral of Eq. (16.15). If $F(s)$ is indeed the optimum filter with fixed Φ's and $F_1(s)$; then by forming the neighboring function $F(s) + \eta(s)$ where $\eta(s)$ is the arbitrary variation and by substituting this neighboring function into Eq. (16.15), we find the first-order variation of the mean-square error as

$$\begin{aligned}\delta\overline{e^2} = \tfrac{1}{2}\int_{-\infty}^{\infty} \eta(-i\omega)\{F(i\omega)[\Phi_{ff}(\omega) + \Phi_{fn}(\omega) + \Phi_{nf}(\omega) + \Phi_{nn}(\omega)] \\ - F_1(i\omega)[\Phi_{ff}(\omega) + \Phi_{nf}(\omega)]\}\, d\omega \\ + \tfrac{1}{2}\int_{-\infty}^{\infty} \eta(i\omega)\{F(-i\omega)[\Phi_{ff}(\omega) + \Phi_{fn}(\omega) + \Phi_{nf}(\omega) + \Phi_{nn}(\omega)] \\ - F_1(-i\omega)[\Phi_{ff}(\omega) + \Phi_{fn}(\omega)]\}\, d\omega \qquad (16.16)\end{aligned}$$

If $F(s)$ is the optimum filter transfer function, then the variation $\delta\overline{e^2}$ should vanish for arbitrary $\eta(s)$. This condition will yield an equation for $F(s)$. However, before we can actually take this step, we have to make some very important modifications of Eq. (16.16).

First, the power spectra $\Phi_{ff}(\omega)$ and $\Phi_{nn}(\omega)$ are symmetrical functions of ω. Thus by taking into account Eq. (16.12), we see that the sum of Φ_{ff}, Φ_{fn}, Φ_{nf}, and Φ_{nn} is an even function of ω. It is thus reasonable to expect that this sum Φ can be "factored" so that

$$\Phi(\omega) = \Phi_{ff}(\omega) + \Phi_{fn}(\omega) + \Phi_{nf}(\omega) + \Phi_{nn}(\omega) = \Psi(i\omega)\Psi(-i\omega) \qquad (16.17)$$

$\Psi(s)$ is by definition a function with poles and zeros in the left-half s plane. $\Psi(-s)$ then is a function with poles and zeros in the right-half s plane. Now the transfer function of the filter, for reasons of stability, can have poles only in the left-half s plane. Thus $F(s)$ and $\eta(s)$ are functions with poles only in the left-half s plane. $F(-s)$ and $\eta(-s)$ are functions with poles only in the right-half s plane. Thus the physical requirements limit the class of functions for $F(s)$ and $\eta(s)$. With this understanding, we can rewrite Eq. (16.16) as

$$\begin{aligned}\delta\overline{e^2} = \tfrac{1}{2}\int_{-\infty}^{\infty} \eta(-i\omega)\Psi(-i\omega)\left[F(i\omega)\Psi(i\omega) - \frac{F_1(i\omega)[\Phi_{ff}(\omega) + \Phi_{nf}(\omega)]}{\Psi(-i\omega)}\right] d\omega \\ + \tfrac{1}{2}\int_{-\infty}^{\infty} \eta(i\omega)\Psi(i\omega)\left[F(-i\omega)\Psi(-i\omega)\right. \\ \left. - \frac{F_1(-i\omega)[\Phi_{ff}(\omega) + \Phi_{fn}(\omega)]}{\Psi(i\omega)}\right] d\omega \qquad (16.18)\end{aligned}$$

However, not all of the second terms in the braces of Eq. (16.18) are important. If $H(s)$ and $K(s)$ are functions with poles in the left-half

s plane, and if for large s they behave like $1/s^n$, with $n \geq 1$, then

$$\int_{-\infty}^{\infty} H(i\omega)K(i\omega)\,d\omega = i \oint H(s)K(s)\,ds$$

where the closed path of integration of the second integral is the imaginary axis and the semicircle enclosing the right-half s plane, as shown in Fig. 16.2a. But the singularities of $H(s)K(s)$ are outside this path, and hence the value of the integral is zero. For the product $H(-i\omega)K(-i\omega)$ with singularities in the right-half s plane, the path of integration can be made to enclose the left-half s plane, as shown in Fig. 16.2b. Then

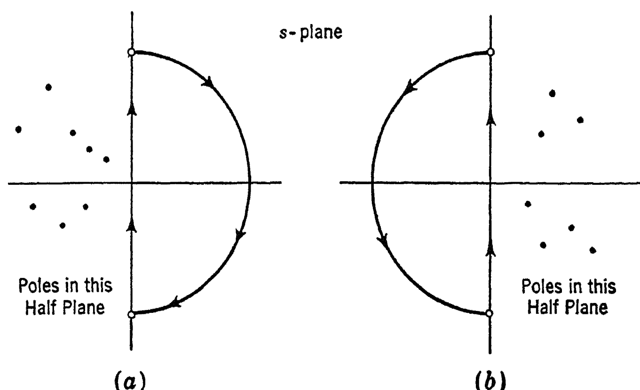


FIG. 16.2

again no singularities will be enclosed, and the value of the integral will be zero. For products $H(i\omega)K(-i\omega)$ or $H(-i\omega)K(i\omega)$, the path will always enclose some singularities, and the integral has value. Thus

$$\int_{-\infty}^{\infty} H(i\omega)K(i\omega)\,d\omega = \int_{-\infty}^{\infty} H(-i\omega)K(-i\omega)\,d\omega = 0 \qquad (16.19)$$

But

$$\left.\begin{aligned} \int_{-\infty}^{\infty} H(i\omega)K(-i\omega)\,d\omega \neq 0 \\ \int_{-\infty}^{\infty} H(-i\omega)K(i\omega)\,d\omega \neq 0 \end{aligned}\right\} \qquad (16.20)$$

With these facts in mind, we divide $F_1(s)\Phi_{ff}(s/i)/\Psi(-s)$ into two parts: one part with singularities in the left-half s plane, denoted by $[\quad]_+$, and another part with singularities in the right-half s plane, denoted by $[\quad]_-$. That is,

$$\frac{F_1(i\omega)\{\Phi_{ff}(\omega) + \Phi_{nf}(\omega)\}}{\Psi(-i\omega)} = \left[\frac{F_1(i\omega)\{\Phi_{ff}(\omega) + \Phi_{nf}(\omega)\}}{\Psi(-i\omega)}\right]_+ + \left[\frac{F_1(i\omega)\{\Phi_{ff}(\omega) + \Phi_{nf}(\omega)\}}{\Psi(-i\omega)}\right]_- \qquad (16.21)$$

Then because of Eqs. (16.19) and (16.20), Eq. (16.18) can be written as

$$\delta\overline{e^2} = \tfrac{1}{2}\int_{-\infty}^{\infty} \eta(-i\omega)\Psi(-i\omega)$$

$$\left\{F(i\omega)\Psi(i\omega) - \left[\frac{F_1(i\omega)\{\Phi_{ff}(\omega) + \Phi_{nf}(\omega)\}}{\Psi(-i\omega)}\right]_+\right\} d\omega + \tfrac{1}{2}\int_{-\infty}^{\infty} \eta(i\omega)\Psi(i\omega)$$

$$\left\{F(-i\omega)\Psi(-i\omega) - \left[\frac{F_1(-i\omega)\{\Phi_{ff}(\omega) + \Phi_{fn}(\omega)\}}{\Psi(i\omega)}\right]_-\right\} d\omega \quad (16.22)$$

Now if $F(s)$ is indeed the transfer function of the optimum filter, then $\delta\overline{e^2}$ should vanish for arbitrary $\eta(s)$. Thus the quantities in the brackets of Eq. (16.22) must vanish. This condition determines the optimum transfer function as

$$F(s) = \frac{1}{\Psi(s)}\left[\frac{F_1(s)\{\Phi_{ff}(s/i) + \Phi_{nf}(s/i)\}}{\Psi(-s)}\right]_+ \quad (16.23)$$

This is the solution of the optimum filter problem given by Wiener and Kolmogoroff. $F(s)$ has poles only in the left-half s plane, since $\Psi(s)$ has zeros only in the left-half s plane. $F(s)$ is thus a stable transfer function. The operation of picking the part of $F_1(s)\{\Phi_{ff}(s/i) + \Phi_{nf}(s/i)\}/\Psi(-s)$ with poles in the left-half s plane can be also done analytically. In fact,

$$F(s) = \frac{1}{2\pi\Psi(s)}\int_0^{\infty} e^{-st}\,dt \int_{-\infty}^{\infty} \frac{F_1(i\omega)\{\Phi_{ff}(\omega) + \Phi_{nf}(\omega)\}}{\Psi(-i\omega)} e^{i\omega t}\,d\omega \quad (16.24)$$

The validity of this equation can be easily verified by contour integration in the complex ω plane. Which of the two equations (16.23) and (16.24) is to be used for actual calculation depends upon the individual cases. Usually, however, Eq. (16.23) is much easier to use. In any event, the properties of the optimum filter are completely determined by the specified operation $F_1(s)$ and the spectra Φ_{ff}, Φ_{fn}, Φ_{nf}, and Φ_{nn} of the signal and noise. When noise is absent, $\Phi_{nn} = \Phi_{fn} = \Phi_{nf} = 0$, and $\Phi_{ff}(\omega) = \Psi(i\omega)\Psi(-i\omega)$, according to Eq. (16.17). Then $F(s) = F_1(s)$ as expected, and the error $e(t)$ is always zero. When there is noise, $F(s)$ is not equal to $F_1(s)$, and the mean-square error cannot be eliminated even with the best filter.

A different interpretation of the operation of picking the part of a function with poles in the left-half s plane can be given. As shown by Eq. (16.2), we can consider $F(i\omega)$ to be the Fourier transform of the response $h(t)$ to a unit impulse. That equation also shows how to calculate $h(t)$ from $F(i\omega)$. Because of the factor $e^{i\omega t}$ in the integrand of Eq. (16.2), the proper paths to take in the complex ω plane are different for positive t and for negative t, and are shown in Fig. 16.3. If $F(s)$ has poles only in the left-half s plane, then $F(i\omega)$ has poles only in the

upper-half ω plane. If $F(s)$ has poles only in the right-half s plane, then $F(i\omega)$ has poles only in the lower-half ω plane. Thus for $t > 0$, the path will enclose poles of $F(i\omega)$ only if $F(s)$ has poles in the left-half s plane, and in that case the response $h(t)$ will be different from zero. For $t < 0$, the path will enclose no pole of $F(i\omega)$ only if $F(s)$ has all its poles in the left-half s plane, and then $h(t) = 0$. On the other hand, if $F(s)$ has poles in the right-half s plane, the case of unstable $F(s)$ according to the present interpretation, $h(t)$ will be different from zero even for negative t. Since the impulse is applied at $t = 0$, response at negative t means response before the impulse is applied. Such behavior is

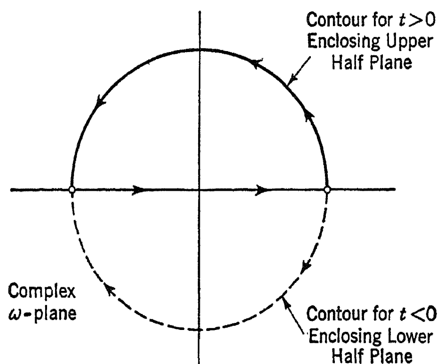


FIG. 16.3

impossible for any physical system. Therefore we can consider the operation of picking the part of function with poles only in the left-half s plane as the operation of making the transfer function *physically realizable*, so that $h(t) = 0$ for $t < 0$. An explanation of the Wiener-Kolmogoroff solution of Eq. (16.22) and (16.23) based upon this concept of a physically realizable transfer function was given by Bode and Shannon.[1]

There remains one point to be cleared. One of the assumptions in the present theory is the possibility of factoring the sum of spectra as shown in Eq. (16.17). This is not always possible for an even positive function $\Phi(\omega)$ of ω. To make factoring possible, $\Phi(\omega)$ must satisfy the *Wiener-Paley criterion*,[2]

$$\int_{-\infty}^{\infty} \frac{|\log \Phi(\omega)|}{1 + \omega^2} \, d\omega < \infty \tag{16.25}$$

Actually $\Phi(\omega)$ is either a constant, as in case of white noise, or approaches zero as $\omega \to \infty$. The Wiener-Paley criterion says then that the approach

[1] H. W. Bode and C. E. Shannon, *Proc. IRE*, **38**, 417–425 (1950).

[2] R. E. A. C. Paley and N. Wiener, "Fourier Transforms in the Complex Domain," Am. Math. Soc. Colloquium Publication, Vol. 19, p. 17, 1934.

to zero at large ω should not be too rapid. An approach of the type ω^{-n} is allowed, but an approach like $e^{-|\omega|}$ or $e^{-\omega^2}$ will make the integral diverge. Thus a $\Phi(\omega)$ of the latter types cannot be factored. Fortunately the spectra of the actually occurring signal and noise are generally ratios of polynomials of ω^2. Factoring according to Eq. (16.17) is thus usually possible.

16.4 Simple Examples. As a simple example, we take the power spectrum of the signal to be

$$\Phi_{ff}(\omega) = \frac{1}{1+\omega^4} \tag{16.26}$$

The noise is assumed to be white noise. Then its power spectrum is flat and a constant; thus

$$\Phi_{nn} = n^4 \tag{16.27}$$

The cross-correlation functions and their Fourier transforms thus vanish, *i.e.*,

$$\Phi_{fn} = \Phi_{nf} = 0 \tag{16.28}$$

Let the problem be to design the optimum filter for differentiating the signal, that is, $F_1(s) = s$. First, we note that

$$\Phi(\omega) = \Phi_{ff}(\omega) + \Phi_{nn}(\omega) = \frac{(1+n^4) + n^4\omega^4}{1+\omega^4}$$

Thus

$$\Phi(s/i) = \Psi(s)\Psi(-s) = \frac{(1+n^4) + n^4 s^4}{1+s^4}$$

Obviously to factor this function is quite straightforward. Remembering that $\Psi(s)$ has zeros and poles only in the left-half s plane, we can immediately write down $\Psi(s)$ as

$$\Psi(s) = \frac{n^2s^2 + n\sqrt{2}\sqrt[4]{1+n^4}\,s + \sqrt{1+n^4}}{s^2+\sqrt{2}\,s+1}$$

Then

$$\begin{aligned}\frac{F_1(s)\Phi_{ff}(s/i)}{\Psi(-s)} &= \frac{s}{(s^2+\sqrt{2}\,s+1)(n^2s^2 - n\sqrt{2}\sqrt[4]{1+n^4}\,s + \sqrt{1+n^4})} \\ &= \frac{As+B}{s^2+\sqrt{2}\,s+1} + \frac{Cs+D}{n^2s^2 - n\sqrt{2}\sqrt[4]{1+n^4}\,s + \sqrt{1+n^4}}\end{aligned}$$

where A, B, C, and D are constants. It is evident that the part with poles only in the left-half s plane is the first term. Hence

$$\left[\frac{F_1(s)\Phi_{ff}(s/i)}{\Psi(-s)}\right]_+ = \frac{As+B}{s^2+\sqrt{2}\,s+1}$$

By solving for A and B, we have, finally,

$$F(s) = \frac{1}{\Psi(s)}\left[\frac{F_1(s)\Phi_{ff}(s/i)}{\Psi(-s)}\right]_+ = \frac{1}{(n^2+\sqrt{1+n^4})(n+\sqrt[4]{1+n^4})} \cdot \frac{(\sqrt[4]{1+n^4}-n)s - n\sqrt{2}}{n^2s^2 + n\sqrt{2}\sqrt[4]{1+n^4}\,s + \sqrt{1+n^4}} \tag{16.29}$$

This is the transfer function for the optimum filter. As the noise is reduced, $n \to 0$, and $F(s)$ approaches s as it should.

Another interesting example is the case of very high noise intensity and a weak signal. Let the noise be white noise and the power spectrum of the noise Φ_{nn} be

$$\Phi_{nn}(\omega) = 1 \tag{16.30}$$

Here again there is no cross correlation between noise and signal, and Eq. (16.28) remains true. Let the power spectrum of the signal be

$$\Phi_{ff}(\omega) = k\varphi(\omega) \tag{16.31}$$

where k is a small quantity and $\varphi(\omega)$ is an even function of ω. If $K(s)$ is the part of $\varphi(s/i)$ with poles only in the left-half s plane, *i.e.*,

$$K(s) = \left[\varphi\left(\frac{s}{i}\right)\right]_+ = \frac{1}{2\pi}\int_0^\infty e^{-st}\,dt \int_{-\infty}^{\infty} \varphi(\omega)e^{i\omega t}\,d\omega \tag{16.32}$$

then since $\varphi(\omega)$ is even in ω,

$$\varphi\left(\frac{s}{i}\right) = K(s) + K(-s) \tag{16.33}$$

and

$$\Phi(\omega) = \Psi(i\omega)\Psi(-i\omega) = 1 + k\varphi(\omega) \approx [1 + kK(i\omega)][1 + kK(-i\omega)] \tag{16.34}$$

Therefore if $F_1(s)$ represents the desired operation on the signal, the optimum filter is given by

$$F(s) \approx \frac{k}{1+kK(s)}\left[\frac{F_1(s)\varphi(s/i)}{1+kK(-s)}\right]_+ \tag{16.35}$$

This is the second approximation for small k. The first approximation is even simpler,

$$F(s) \approx k\left[F_1(s)\varphi\left(\frac{s}{i}\right)\right]_+ \tag{16.36}$$

As a specific example, let the power spectrum of the signal be specified by

$$\varphi(\omega) = \frac{1}{1+\omega^4}$$

and let $F_1(s) = s$. Then when k is small,

$$F(s) \approx -\frac{k}{2\sqrt{2}} \frac{1}{s^2 + \sqrt{2}\,s + 1}$$

This result checks with Eq. (16.29) by making n very large and $k = 1/n^4$. Thus under strong noise interference, the optimum differentiating transfer function is very much distorted and bears no resemblance at all to $F_1(s) = s$.

16.5 Applications of Wiener-Kolmogoroff Theory. Besides the simple examples discussed in the preceding section, there are a number of very important applications of the Wiener-Kolmogoroff theory. We shall consider several of them in this section.

Predicting Filters. These filters take the input $x(t)$, the sum of signal $f(t)$ and noise $n(t)$, and give an output $y(t)$ which is the closest representation of the signal not at t but at $t + \alpha$, where α is positive. Thus

$$F_1(s) = e^{\alpha s} \tag{16.37}$$

Now let us assume that the signal is a random switching function. Then according to Eq. (9.50), the power spectrum can be written as

$$\Phi_{ff} = \frac{1}{1 + \omega^2} \tag{16.38}$$

The noise, assumed to be white noise, has a power spectrum

$$\Phi_{nn} = n^2 \tag{16.39}$$

The cross spectra vanish. Then

$$\Phi(\omega) = \Psi(i\omega)\Psi(-i\omega) = \frac{(1 + n^2) + n^2\omega^2}{1 + \omega^2}$$

Therefore

$$\Psi(i\omega) = \frac{\sqrt{1 + n^2} + ni\omega}{1 + i\omega}$$

Hence

$$\frac{F_1(i\omega)\Phi_{ff}(\omega)}{\Psi(-i\omega)} = \frac{e^{i\alpha\omega}}{(1 + i\omega)(\sqrt{1 + n^2} - ni\omega)}$$

In this case, it will be necessary to use Eq. (16.24) for calculating $F(s)$. First of all, for $t > 0$,

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{F_1(i\omega)\Phi_{ff}(\omega)}{\Psi(-i\omega)} e^{i\omega t}\, d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{i\omega(t+\alpha)}\, d\omega}{(1 + i\omega)(\sqrt{1 + n^2} - ni\omega)}$$

$$= \frac{e^{-(t+\alpha)}}{n + \sqrt{1 + n^2}}$$

Hence, according to Eq. (16.24),

$$F(s) = \frac{1+s}{\sqrt{1+n^2}+ns}\int_0^\infty e^{-st}\frac{e^{-(t+\alpha)}}{n+\sqrt{1+n^2}}\,dt$$

$$= \frac{(1+s)e^{-\alpha}\int_0^\infty e^{-(s+1)t}\,dt}{(n+\sqrt{1+n^2})(\sqrt{1+n^2}+ns)}$$

Therefore, finally, the optimum predicting filter of predicting time α is characterized by the transfer function $F(s)$,

$$F(s) = \frac{e^{-\alpha}}{(n+\sqrt{1+n^2})}\,\frac{1}{(\sqrt{1+n^2}+ns)} \tag{16.40}$$

Lagging Filters. This type is similar to the predicting filters except that now the "predicting" time α is negative. A straightforward adaptation of the preceding calculation to this case will not lead to success.

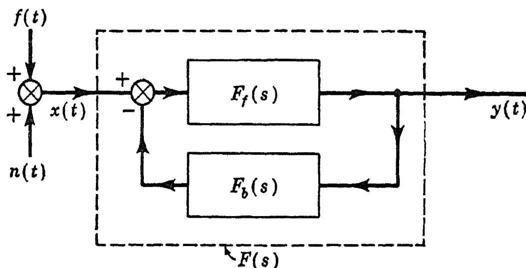


FIG. 16.4

In fact there is no linear system of finite order that will exactly fulfill the role of an optimum lagging filter. A more direct solution can be obtained by using the approximation

$$F_1(s) = e^{\alpha s} \approx \left[\frac{1+(\alpha s/2\nu)}{1-(\alpha s/2\nu)}\right]^\nu \qquad \begin{array}{l}\alpha < 0\\ \nu = \text{integer}\end{array} \tag{16.41}$$

We then obtain an approximation to the optimum lagging filter.

Servomechanisms with Noise. Let $F_f(s)$ be the transfer function of the forward circuit and $F_b(s)$ be the transfer function of the feedback circuit of a feedback servomechanism, as shown in Fig. 16.4. Let $F_1(s)$ represent the desired operation on the signal. The problem is to find the optimum $F_b(s)$ with $F_f(s)$, $F_1(s)$, and the signal and noise properties specified. As shown in the figure, the equivalent $F(s)$ is

$$F(s) = \frac{F_f(s)}{1+F_f(s)F_b(s)}$$

But according to the theory, the optimum $F(s)$ is given by Eq. (16.23) or Eq. (16.24). Knowing $F(s)$, we can obtain the optimum transfer

function $F_b(s)$ for the feedback circuit as

$$F_b(s) = \frac{1}{F(s)} - \frac{1}{F_f(s)} \tag{16.42}$$

Noisy Servomechanisms. In the previous paragraph we assumed the source of noise to be outside the servomechanism and the servo itself quiet. However, very often there is noise generated in the servomechanism. For instance, the system shown in Fig. 16.5 has the external noise $n(t)$, and in addition an internal noise $m(t)$ from the output measuring instrument, or an output disturbance. Here again, let $F_f(s)$ be the transfer function of the forward circuit, $F_b(s)$ be the transfer function

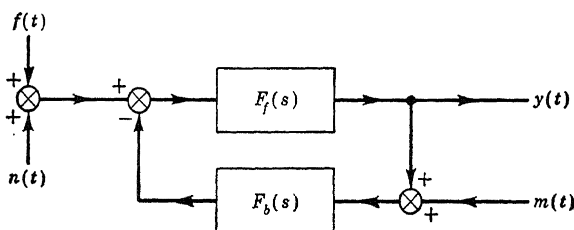


Fig. 16.5

of the feedback circuit, and $F_1(s)$ be the desired operation on the signal. Let the Laplace transforms of $f(t)$, $n(t)$, $m(t)$, $y(t)$, and $z(t)$ be $S(s)$, $N(s)$, $M(s)$, $Y(s)$, and $Z(s)$, respectively. Then

$$F_f(s)\{S(s) + N(s) - F_b(s)[Y(s) + M(s)]\} = Y(s)$$

and

$$Z(s) = F_1(s)S(s)$$

Then the Laplace transform $E(s)$ of the error $e(t)$ is

$$E(s) = Y(s) - Z(s) = \frac{F_f(s)}{1 + F_f(s)F_b(s)}[S(s) + N(s)] - \frac{F_f(s)F_b(s)}{1 + F_f(s)F_b(s)} M(s) - F_1(s)S(s)$$

Let us put

$$F(s) = \frac{F_f(s)F_b(s)}{1 + F_f(s)F_b(s)} \tag{16.43}$$

Then we have

$$1 - F(s) = \frac{1}{1 + F_f(s)F_b(s)}$$

and

$$E(s) = [1 - F(s)][F_f(s)S(s) + F_f(s)N(s) - F_1(s)S(s)] - F(s)[M(s) + F_1(s)S(s)]$$

This equation indicates that our servo problem is equivalent to the filter problem, with the filter transfer function $F(s)$, and the equivalent signal

input $S'(s)$ and equivalent noise input $N'(s)$ given by

$$\left.\begin{aligned} S'(s) &= \{F_f(s) - F_1(s)\}S(s) + F_f(s)N(s) \\ N'(s) &= M(s) + F_1(s)S(s) \end{aligned}\right\} \tag{16.44}$$

The original problem of optimum $F_b(s)$ is now reduced to that of optimum $F(s)$, with the equivalent signal and noise independent of the unknown $F_b(s)$.

Let us assume that the original signal and noises are independent and only autocorrelations exist. Then we have only the power spectra Φ_{ff}, Φ_{nn}, and Φ_{mm}. By using Eq. (16.44), the Φ's of the equivalent filter problem are found,

$$\left.\begin{aligned} \Phi_{f'f'}\left(\frac{s}{i}\right) &= [F_f(s) - F_1(s)][F_f(-s) - F_1(-s)]\Phi_{ff}\left(\frac{s}{i}\right) \\ &\qquad + F_f(s)F_f(-s)\Phi_{nn}\left(\frac{s}{i}\right) \\ \Phi_{f'n'}\left(\frac{s}{i}\right) &= [F_f(-s) - F_1(-s)]F_1(s)\Phi_{ff}\left(\frac{s}{i}\right) \\ \Phi_{n'f'}\left(\frac{s}{i}\right) &= F_1(-s)[F_f(s) - F_1(s)]\Phi_{ff}\left(\frac{s}{i}\right) \\ \Phi_{n'n'}\left(\frac{s}{i}\right) &= \Phi_{mm}\left(\frac{s}{i}\right) + F_1(s)F_1(-s)\Phi_{ff}\left(\frac{s}{i}\right) \end{aligned}\right\} \tag{16.45}$$

Thus the equivalent filter problem has cross spectra $\Phi_{f'n'}$ and $\Phi_{n'f'}$, in spite of the uncorrelated signal and noises in the original problem. The function to be factored is then, using Eq. (16.45),

$$\Phi(\omega) = \Psi(i\omega)\Psi(-i\omega) = F_f(i\omega)F_f(-i\omega)\{\Phi_{ff}(\omega) + \Phi_{nn}(\omega)\} + \Phi_{mm}(\omega) \tag{16.46}$$

The optimum $F(s)$ is, according to Eq. (16.23),

$$F(s) = \frac{1}{\Psi(s)} \left[\frac{F_f(s)F_f(-s)\{\Phi_{ff}(s/i) + \Phi_{nn}(s/i)\} - F_1(s)F_f(-s)\Phi_{ff}(s/i)}{\Psi(-s)}\right]_+ \tag{16.47}$$

When $F(s)$ is known, Eq. (16.43) gives the optimum transfer function $F_b(s)$ for the feedback circuit as

$$F_b(s) = \frac{1/F_f(s)}{[1/F(s)] - 1} \tag{16.48}$$

Saturation Constraint. Consider the servomechanism of Fig. 16.6, designed to make the output $y(t)$ follow as closely as possible the input $f(t) = x(t)$. The transfer function of the amplifier is $F_a(s)$, and the

transfer function of the servo motor is $F_m(s)$. If the design condition is to make the mean square of the error $e(t) = y(t) - f(t)$ as small as possible by varying $F_a(s)$, then the power of the control input to the servo motor may reach a very high level during operation. To avoid this, we must specify that the average of the power of the input to the motor must be a specified value. The mean-square error in this case is

$$\overline{e^2} = \tfrac{1}{2}\int_{-\infty}^{\infty}\left[\frac{F_a(i\omega)F_m(i\omega)}{1+F_a(i\omega)F_m(i\omega)} - 1\right]\left[\frac{F_a(-i\omega)F_m(-i\omega)}{1+F_a(-i\omega)F_m(-i\omega)} - 1\right]\Phi_{ff}(\omega)\,d\omega \quad (16.49)$$

The average power of the input to the servo motor is represented by the mean square of the servo motor signal. This is to be kept at the fixed

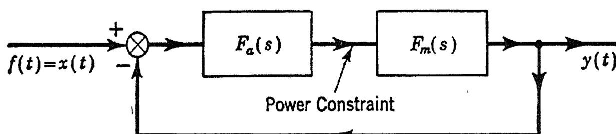


FIG. 16.6

value σ^2. Then

$$\sigma^2 = \tfrac{1}{2}\int_{-\infty}^{\infty}\left|\frac{F_a(i\omega)}{1+F_a(i\omega)F_m(i\omega)}\right|^2 \Phi_{ff}(\omega)\,d\omega \quad (16.50)$$

By using the Lagrange-multiplier method, this problem of minimizing $\overline{e^2}$ with the constraint of Eq. (16.50) can be converted into minimizing $\overline{e^2} + \lambda\sigma^2$, λ being the constant amplifier. Therefore the integral to be minimized is

$$\overline{e^2} + \lambda\sigma^2 = \tfrac{1}{2}\int_{-\infty}^{\infty}\left\{[F(i\omega) - 1][F(-i\omega) - 1]\Phi_{ff}(\omega) + F(i\omega)F(-i\omega)\frac{\lambda\Phi_{ff}(\omega)}{F_m(i\omega)F_m(-i\omega)}\right\}d\omega \quad (16.51)$$

where

$$F(s) = \frac{F_a(s)F_m(s)}{1+F_a(s)F_m(s)} \quad (16.52)$$

Comparing Eq. (16.51) with the integral of Eq. (16.15), we see that the present problem is equivalent to the filter problem with $F_1(s) = 1$, $\Phi_{fn} = \Phi_{nf} = 0$, and the equivalent noise power spectrum

$$\Phi_{nn}(\omega) = \frac{\lambda\Phi_{ff}(\omega)}{F_m(i\omega)F_m(-i\omega)}$$

The function $\Phi(\omega)$ to be factored is thus

$$\Phi(\omega) = \Psi(i\omega)\Psi(-i\omega) = \left[1 + \frac{\lambda}{F_m(i\omega)F_m(-i\omega)}\right]\Phi_{ff}(\omega) \qquad (16.53)$$

According to Eq. (16.23), the optimum $F(s)$ is given by

$$F(s) = \frac{1}{\Psi(s)}\left[\frac{\Phi_{ff}(s/i)}{\Psi(-s)}\right]_+ \qquad (16.54)$$

Equations (16.52) and (16.54) determine the transfer function $F_a(s)$ of the optimum amplifier, with the exception of the constant λ. This constant is fixed by the power level σ^2, according to Eq. (16.50).

The discussion in the preceding paragraphs has perhaps demonstrated the wide range of problems that can be solved by the Wiener-Kolmogoroff theory. In fact, the problem of servomechanism design for random input, formulated in Sec. 9.12, can also be conveniently solved by this theory. Such an application to servomechanisms is another example of design according to specified criteria, a subject treated in its generality in Chap. 14. Whenever such criteria can be formulated, the optimum system characteristics are completely determined. Furthermore, because of the particular type of criteria chosen and the properties of the controlled system, the resultant control system is linear and has constant coefficients. This concept of a more specific design of servomechanisms is thus one step beyond the design principles of feedback servos discussed in the earlier chapters. Boksenbom and Novik[1] were perhaps the first to suggest this particular application of filter theory.

16.6 Optimum Detecting Filter. In many control systems, the problem is the detection of the signal $f(t)$ under heavy noise interference. In this problem, the pattern of the signal is generally known; what has to be detected is the time instant t_0 when the signal has reached the expected value $f(t_0)$. For instance, in the case of radar, we know that the signal is a pulse having a specified shape. The problem is to know when the signal has reached its maximum strength. If the signal has its maximum at the instant t_0, then the filter should give the least distortion at the instant t_0. Therefore the optimum filter must be so designed that the value of the signal after filter action, $f_o(t)$, at the instant t_0 is, in fact, $f(t_0)$. Thus the constraint is

$$f_o(t_0) = f(t_0) = \text{a const.} \qquad (16.55)$$

The noise input to the filter is $n(t)$; the corresponding output is $n_o(t)$. We wish to reduce the noise as much as possible by filtering, and thus we can write

$$\overline{n_o^2(t)} = \text{min.} \qquad (16.56)$$

[1] A. S. Boksenbom, D. Novik, *NACA TN* 2939 (1953).

The problem is to determine the transfer function $F(s)$ of the filter or its equivalent, the response $h(t)$ of the filter to a unit impulse, with given $f(t)$ and t_0, and the noise characteristics R_{nn} or Φ_{nn}, such that

$$\overline{n_o^2(t)} - 2\lambda f_o(t_0) = \min. \tag{16.57}$$

where λ is the Lagrange multiplier. The solution of this problem is given in this generalized form by Zadeh and Ragazzini.[1] We shall follow their analysis.

If $\Phi_{nn}(\omega)$ can be factored, so that

$$\Phi_{nn}(\omega) = \Psi(i\omega)\Psi(-i\omega) \tag{16.58}$$

where $\Psi(s)$ has zeros and poles only in the left-half s plane; then it is convenient to consider two amplifiers in series, where one has the transfer

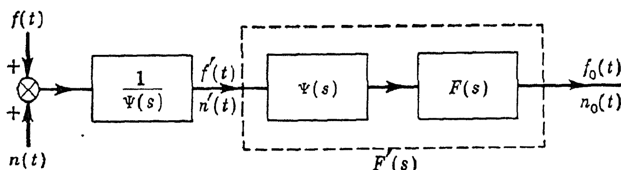


FIG. 16.7

function $1/\Psi(s)$ and the other the transfer function $\Psi(s)$, both in series again with the filter, as shown in Fig. 16.7. This system is equivalent to the original system. We consider the last two transfer functions as one unit $F'(s)$, *i.e.*,

$$F'(s) = \Psi(s)F(s) \tag{16.59}$$

with the response $h'(t)$ to an impulse. The inputs to $F'(s)$ are the signal $f'(t)$ and the noise $n'(t)$. If $S(i\omega)$ is the Fourier transform of the signal $f(t)$, *i.e.*,

$$S(i\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt \tag{16.60}$$

then

$$f'(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{S(i\omega)}{\Psi(i\omega)}\,e^{i\omega t}\,d\omega \tag{16.61}$$

The power spectrum of $n'(t)$ is now, according to Eq. (9.71),

$$\frac{\Phi_{nn}(\omega)}{\Psi(i\omega)\Psi(-i\omega)} = 1$$

Therefore the noise is now a white noise with the autocorrelation function

$$R_{n'n'}(\tau) = \delta(\tau) \tag{16.62}$$

[1] L. A. Zadeh, J. R. Ragazzini, *Proc. IRE*, **40**, 1223–1231 (1952).

The outputs from the system can now be written as

$$f_o(t) = \int_0^\infty h'(\tau)f'(t-\tau)\,d\tau \tag{16.63}$$

and

$$n_o(t) = \int_0^\infty h'(\tau)n'(t-\tau)\,d\tau. \tag{16.64}$$

The mean square of output noise is then the assembly average of $n_o^2(t)$. Or

$$\begin{aligned}\overline{n_o^2(t)} &= \int_0^\infty d\tau \int_0^\infty d\tau' h'(\tau)h'(\tau')\overline{n'(t-\tau)n'(t-\tau')} \\ &= \int_0^\infty \int_0^\infty h'(\tau)h'(\tau')R_{n'n'}(\tau-\tau')\,d\tau\,d\tau'\end{aligned}$$

But the correlation of noise $n'(t)$ is given by Eq. (16.62). Hence

$$\overline{n_o^2} = \int_0^\infty [h'(t)]^2\,dt \tag{16.65}$$

Therefore the minimization condition of Eq. (16.57) can be converted by Eqs. (16.63) and (16.65) to

$$\int_0^\infty [h'(t)]^2\,dt - 2\lambda \int_0^\infty h'(t)f'(t_0-t)\,dt = \text{min.}$$

This can be rewritten as

$$\int_0^\infty [h'(t) - \lambda f'(t_0-t)]^2\,dt - \lambda^2 \int_0^\infty [f'(t_0-t)]^2\,dt = \text{min.} \tag{16.66}$$

But with $f(t)$ and $\Phi_{nn}(\omega)$ specified by the problem, $f'(t)$ is a fixed function calculated by Eq. (16.61). Therefore the second integral of Eq. (16.66) is a fixed constant. The first integral is positive or zero. For the minimum of the sum, the first integral of Eq. (16.66) should be zero. This is so only if the quantity within the square bracket of that integral vanishes. Thus

$$h'(t) = \lambda f'(t_0-t) \qquad \text{for } t \geq 0 \tag{16.67}$$

Naturally, for physically realizable systems $h'(t) \equiv 0$ for $t < 0$. In other words, the optimum response of F' to an impulse is identical, for positive t, with the image of $f'(t)$ with respect to $t_0/2$. This result for white noise has been known for some time and was first derived by D. O. North.

With the result of Eq. (16.67), the optimum filter $F(s)$ of the original problem can be immediately written down with the help of Eqs. (16.59) and (16.61),

$$F(s) = \frac{\lambda}{2\pi\Psi(s)} \int_0^\infty dt\, e^{-st} \int_{-\infty}^\infty \frac{S(-i\omega)}{\Psi(-i\omega)} e^{i\omega(t-t_0)}\,d\omega \tag{16.68}$$

where λ is to be finally fixed by the constant $f(t_0)$ in Eq. (16.55). This result is very similar to that of Wiener's optimum filter given by Eq. (16.24). In fact we can write

$$F(s) = \frac{\lambda}{\Psi(s)} \left[\frac{e^{-st_0} S(-s)}{\Psi(-s)} \right]_+ \tag{16.69}$$

where $[\ \ ]_+$ again denotes picking the part of the function within the bracket having poles only in the left-half s plane, or making the transfer function physically realizable. Equations (16.68) and (16.69) are the formulae for optimum detecting filters given by Zadeh and Ragazzini. The derivation is made elementary by using the artifice of an equivalent problem (Fig. 16.7), in which the noise component of the input is a white noise. This device is generally very useful in simplifying the analysis of complex optimization problems.

16.7 Other Optimum Filters. One of the basic assumptions in the filter theory discussed in the preceding sections is the stationarity of the random functions, either as signal or as noise. No true stationarity can hold for very long time, because of the natural drift of the system or a purposeful change in the state of operation. More likely, the random inputs are stationary only for a limited time interval T. For time intervals longer than T, the random functions are not stationary. Therefore, if the filter is designed with the assumption of a stationary random function and if the characteristic time of the filter is much larger than T, reality will not correspond to the theory, and the performance of the filter suffers. In such a case, it would be better to deviate from the theoretical "optimum" filter and use one with a shorter characteristic time.

A more satisfactory solution is to include the time T explicitly in our theory. We can do this by requiring the impulse response $h(t)$ of the filter to be zero outside the interval $0 \leq t \leq T$. Then the output $y(t)$ is calculated from the input $x(t)$ as follows:

$$y(t) = \int_0^T h(\tau)x(t - \tau)\, d\tau \tag{16.70}$$

This equation demonstrates the fact that the output is dependent upon the input only for a finite time T back from the present. Therefore such filters may be called filters of *finite memory*. The filters discussed in the previous sections, having $T \to \infty$, are thus *infinite-memory filters*.

Finite memory filters are discussed by Zadeh and Ragazzini. They[1] give a solution of the optimum filter for the problem of detection similar to that of the preceding section. They[2] also give a solution of the

[1] Zadeh and Ragazzini, *op. cit.*

[2] L. A. Zadeh, J. R. Ragazzini, *J. Appl. Phys.*, **21**, 645–654 (1950).

optimum filter for a more complicated problem with the signal input composed of two parts, a stationary random function and a nonrandom function expressible in a polynomial of nth degree in t. For this problem of a finite-memory filter, the performance criteria are, first, vanishing mean error, and second, minimum mean-square error. The vanishing of the mean error is no longer automatically ensured because of the nonrandom part of the signal. The solutions given by their theory for these problems are, however, generally difficult to realize by simple RC circuits. In fact, even for the simpler problem of infinite-memory filters, the solution specified by Eqs. (16.68) and (16.69) is sometimes difficult to build. Practical filters can be only approximations to the theoretical optimum. Then the value of the theoretical solution lies mainly in giving a guide for design and a standard of ideal performance.

16.8 General Filtering Problem. Of course, it will be still possible to use the complicated theoretical optimum filter design if we abandon the inadequate RC circuits and use an analog computer or even a digital computer to serve as the filter. Then the theoretical optimum performance can actually be attained. However, the introduction of an electromechanical computer as a component of the filtering system greatly increases the complexity of the over-all system and can be justified only in very critical cases. If we have made the system very complicated at high cost, we may ask whether we have actually obtained the very best performance. The optimum performance in the theory discussed in the previous sections is only optimum within the limitations of the assumptions of the theory. For instance, two random signals with the same correlation function or the same power spectrum, according to the theory developed, require the same optimum filter. This is, in a sense, a certain looseness in design criteria. Surely, if we have more statistical information about the signal than just the power spectrum, we should be able to distinguish these two signals and to improve our design by utilizing such additional knowledge. Then we can obtain even better performance than possible with the so-called "optimum" filter. It is evident that this generalized approach to the filtering problem must require a more advanced theory of probability than we have used. The recently developed science of *information theory* may also find important applications here. A beginning[1] has been made in this "probabilistic" approach to the problem of detecting a signal in noise. But much remains to be done.

The Wiener-Kolmogoroff theory of the optimum filter is based upon the mean-square-error criterion. By using this criterion, we essentially

[1] See for instance P. M. Woodward, I. L. Davies, *Phil. Mag.*, **41**, 1001–1017 (1950); *Proc. IRE*, **39**, 1521–1524 (1951); *J. Inst. Elec. Engrs. London*, **99**(3), 37–51 (1952); and T. G. Slattery, *Proc. IRE*, **40**, 1232–1236 (1952).

put the emphasis on minimizing the large errors, without much consideration of the small errors. However, on many occasions, we may be most interested in making the frequent error as small as possible, while not being particular about an infrequent large error. It is also possible that the probability function is very lopsided, with the mean far from the mode. For such cases, the mean-square-error criterion is entirely inappropriate. As a simple example, we consider the problem of predicting whether tomorrow will be a clear day, quoted by Bode and Shannon.[1] Since clear days are in the majority, and there are no days with negative precipitation to balance days when there is precipitation, the probability function is very much lopsided. With this function, the average point, which is the one given by a prediction minimizing the mean-square error, might be represented by a day with a light drizzle. To a man planning a picnic, however, such a prediction would have no value at all. He is interested in the probability of having a really clear day, since even a small amount of precipitation would ruin a picnic.

The theory of the optimum filter of the present chapter also assumes a linear filter in that the differential equation relating the input to the output of the filter is a linear equation with constant coefficients. This is clearly a self-imposed limitation, taken with the purpose of simplifying the theory and with the knowledge that such filters are easily synthesized out of *RC* circuits. For controlled systems that have time-varying characteristics, such as the ballistic guidance problem of Chap. 13, such filters are clearly not suitable. In this case of time-varying systems, the proper filter should also have time-varying characteristics. If the filter is still linear, so that the principle of superposition holds, the input-output relationship is still controlled by the impulse response. But now this response is a function h of two time variables t and t^*. t is the time instant when the response is taken, t^* is the time instant when the impulse is applied. Then the output $y(t)$ can be calculated from the input $x(t)$ as follows:

$$y(t) = \int_0^\infty h(t,t-\tau)x(t-\tau)\,d\tau \tag{16.71}$$

The optimum filter problem is then that of first determining $h(\tau,t)$ and then of finding a way to actually implement this optimum response to an impulse.

[1] Bode and Shannon, *op. cit.*

CHAPTER 17

ULTRASTABILITY AND MULTISTABILITY

In the preceding chapters, we have indicated how control systems of great complexity can be designed to give almost any specified performance. Of course, the greater the complexity of the system, the greater is the chance of malfunction caused by mistakes in assembling the system or the failure of an individual component of the system. Therefore, for complex control systems, the problem of the reliability of the design in actual operation becomes one of utmost importance. In this and the last chapter, we shall consider this problem from two different points of view.

In this chapter we shall discuss the possibility of building into the system a certain measure of flexibility and adaptability so that incidental and unexpected mistakes in design will be automatically corrected by the control system itself without external human aid. Such a control system is thus capable of *learning* how to behave properly and has almost the *homeostatic mechanisms* of living organisms, which enable them to survive under varying conditions of environment. Naturally this concept of self-adjustments in a complicated system came from the study of the behavior of living creatures, because there these characteristics are most evident. In fact, our discussions in this chapter are based upon a remarkable book by W. R. Ashby[1] on the origin of the nervous system's unique ability to produce adaptive behavior. There may be different opinions about how the brain of an animal is actually constructed. But our task is simply to indicate that it is possible to achieve such adaptive behavior by mechanical means. Whether the suggested mechanism is the only one possible does not concern us here.

17.1 Ultrastable System. For simplicity, let us consider an *autonomous system* specified by two variables y_1 and y_2, such as those considered in Sec. 10.5. The phase plane is then the y_1y_2 plane. If t is the time, then the simultaneous differential equations determining the behavior of the system can be written as

$$\left.\begin{aligned} \frac{dy_1}{dt} &= f_1(y_1,y_2;\zeta) \\ \frac{dy_2}{dt} &= f_2(y_1,y_2;\zeta) \end{aligned}\right\} \qquad (17.1)$$

[1] W. R. Ashby, "Design for a Brain," John Wiley & Sons, Inc., New York, 1952.

where we have included in the functions f_1 and f_2 an extra parameter ζ, indicating that the functional relationships between dy_1/dt and dy_2/dt on one hand, and y_1 and y_2 on the other are fixed only if the parameter ζ is fixed. In particular, we shall allow ζ to take a series of discrete values. The pattern of behavior of the system is determined by lines of the loci of the point (y_1,y_2) as time increases, starting from various initial points in the phase plane. Clearly then, there are as many different patterns of behavior of the system as there are different values of the discrete parameter ζ. For instance, the linear system discussed in Sec. 10.5 (with $y_1 = y$ and $y_2 = \dot{y}$) has the parameter ζ. If ζ is capable of taking five different values, one negative, less than -1; one negative, between -1 and 0; one equal to 0; one positive, between 0 and 1; and finally one positive, greater than 1, then the five patterns of behavior of the system are indicated by Figs. 10.12 to 10.16. Another example would be

$$\left.\begin{aligned}\frac{dy_1}{dt} &= a_{11}(\zeta)y_1 + a_{12}(\zeta)y_2\\ \frac{dy_2}{dt} &= a_{21}(\zeta)y_1 + a_{22}(\zeta)y_2\end{aligned}\right\} \tag{17.2}$$

where the coefficients a_{11}, a_{12}, a_{21}, and a_{22} are monotonic functions of ζ. Then there are as many different sets of these coefficients as there are different values of the parameter ζ. Each set of the coefficients gives a definite pattern of behavior.

Some of the patterns of behavior will be stable in that all lines of behavior tend to a point in the phase plane, the stable equilibrium point. Some of the patterns of behavior will be unstable in that the lines of behavior tend to diverge from the equilibrium point. Satisfactory performance of the system naturally requires stability. We shall achieve the desired adaptive behavior of the system if we can cause the system to reject automatically the unstable patterns of behavior and to retain the stable patterns of behavior. Now see what happens if we surround the desired equilibrium point of the system by a closed boundary and if we build the system with a switching device such that the parameter ζ will jump to a different value whenever the line of behavior of the system crosses the boundary. If we start with a pattern of behavior as shown in Fig. 17.1*a* at the point P_0, the pattern of behavior is unstable, and the system will cross the boundary at the point P_1. At the instant of crossing, the switching device acts so that the parameter ζ jumps to a different value, and the pattern of behavior of the system changes to that of Fig. 17.1*b*. This pattern is also unstable, and the system moves from point P_1 to P_2, where it crosses the boundary again. The switching device changes ζ, and the pattern of behavior becomes that of Fig. 17.1*c*. But this pattern of behavior, although it contains a stable equilibrium

point, requires the system to move out of the boundary at P_2. Thus the switching device operates a third time and changes the pattern of behavior to that of Fig. 17.1*d*. This is the stable pattern, and the system moves from the point P_2 to the equilibrium point P_3. This pattern will be retained because, in a condition of stability, the system will not cross the boundary and thus will not activate the switching device to change ζ.

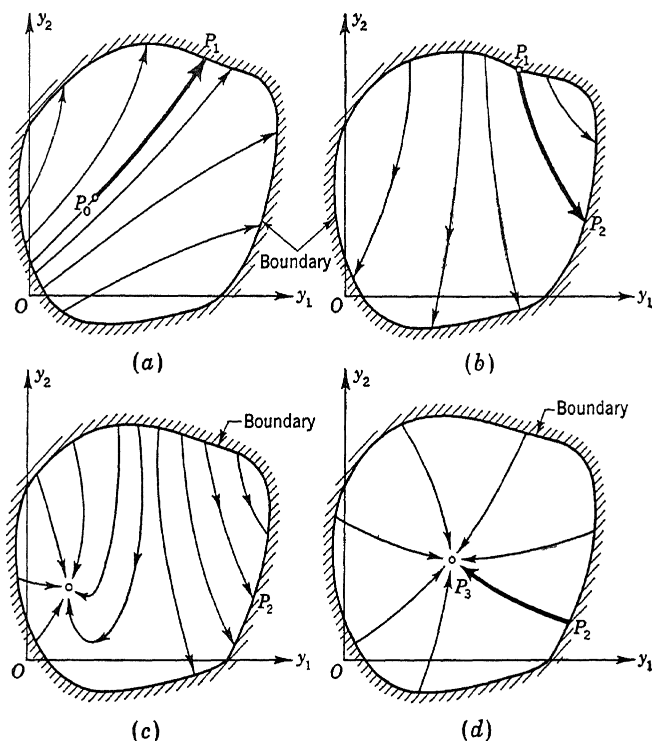


FIG. 17.1

Therefore, with only the addition of the switching device and a prescribed *switching boundary* in the phase plane, the system is made to seek stability automatically, rejecting unstable patterns and retaining stable patterns of behavior. Furthermore, the switching action in changing the parameter ζ can be entirely random. If the stable pattern is attained by the first switching, so much the better. But the end result is the same, whether the system has to switch once or three times. Stability is always achieved. Therefore we are able to obtain *purposeful, goal-seeking* behavior by purely mechanical means. Such a system is thus automatically stable; it does not have to be designed to have stabil-

ity, it will become stable by learning. Therefore it has something more than just stability and will be called an *ultrastable system,* after Ashby.

Our two-variable autonomous system can be generalized to an n-variable system with the variables y_i, where $i = 1, \ldots, n$. Then the differential equations can be written as

$$\frac{dy_i}{dt} = f_i(y_1, y_2, \ldots, y_i, \ldots, y_n; \zeta) \qquad \text{for } i = 1, \ldots, n \tag{17.3}$$

where ζ is the parameter. ζ jumps to a different value whenever the line of behavior in the phase space y_i crosses the switching boundary. The switching boundary here is a closed hypersurface of $n - 1$ dimensions in the phase space of n dimensions. Such a system is also ultrastable.

17.2 An Example of an Ultrastable System. In order to demonstrate the behavior of an ultrastable system, Ashby constructed a relatively simple system of four variables, which he called the homeostat.[1] The four variables y_1, y_2, y_3, and y_4 are the angular deflections of four magnets whose motion is heavily damped. The position of each magnet is controlled by four coils, each of which is fed with currents generated by the angular position of the four magnets. Because of the heavy damping, the motion of the magnets is slow, and the inertial forces can be neglected. The equations of motion can be obtained by equating the turning torques produced by the coils to the damping torques. Thus

$$\left.\begin{aligned}
\frac{dy_1}{dt} &= a_{11}y_1 + a_{12}y_2 + a_{13}y_3 + a_{14}y_4 \\
\frac{dy_2}{dt} &= a_{21}y_1 + a_{22}y_2 + a_{23}y_3 + a_{24}y_4 \\
\frac{dy_3}{dt} &= a_{31}y_1 + a_{32}y_2 + a_{33}y_3 + a_{34}y_4 \\
\frac{dy_4}{dt} &= a_{41}y_1 + a_{42}y_2 + a_{43}y_3 + a_{44}y_4
\end{aligned}\right\} \tag{17.4}$$

The magnitude of the coefficients a can be modified by the experimenter by regulating the current in the coils with a variable potentiometer. The sign of the a's can also be changed by the commutators in the coil circuits. In addition, for each magnet, one of the coils is fed by current passing through a uniselector which has 25 possible positions. The position of the uniselector will jump randomly to a different one whenever the angular deflection of that magnet reaches 45 degrees in either direction. Thus four of the coefficients a_{ij}, for $i, j = 1, 2, 3, 4,$ are each capable of having one of the 25 values chosen randomly by the four uniselectors. For each magnet, or each variable y_i, we have a block diagram similar to that shown in Fig. 17.2.

[1] W. R. Ashby, *Electronic Eng.*, **20**, 379 (1948).

The switching boundary of the homeostat is thus a "cube" in the four-dimensional phase space centered around the origin with sides corresponding to 90-degree angular motion. For every setting of the a's by the experimenter, the four uniselectors give $25^4 = 390{,}625$ combinations of the four coefficients controlled by them. Therefore, for every hand setting, there are 390,625 patterns of behavior of the homeostat, some

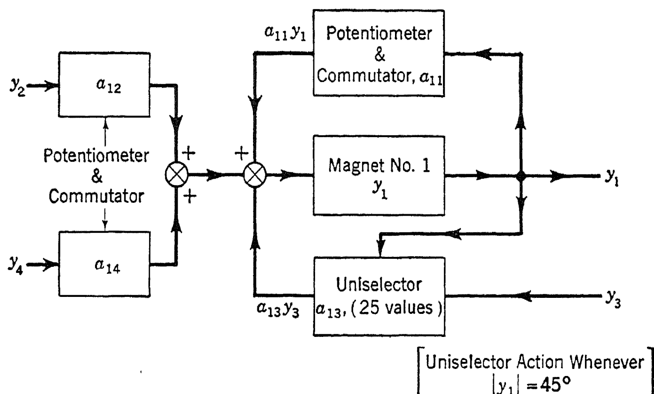


FIG. 17.2

stable, some unstable. The unstable patterns will, however, be automatically rejected.

The ultrastability can now be demonstrated. First, for simplicity, a single unit is shown arranged to feed back into itself through a single uniselector; the other coils are disconnected. The behavior of the single magnet is shown in Fig. 17.3, where the upper curve is the deflection

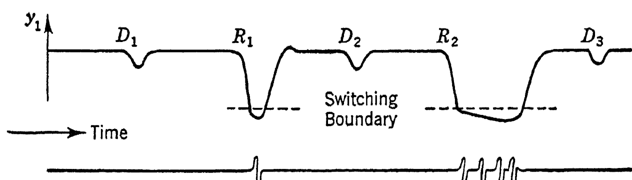


FIG. 17.3

of the magnet and the lower curve indicates the uniselector action. At D_1, the magnet is disturbed by hand; but the uniselector position happens to give a stable pattern of behavior, and the deflection is promptly corrected. At R_1, the feedback is reversed by hand. The old uniselector position now makes the system unstable, and the deflection of the magnet reaches the switching boundary (the dotted line in the figure). The uniselector acts. After one jump of the uniselector, the pattern becomes stable, as shown by the test disturbance at D_2. At R_2, the feedback is again reversed by hand. This time it takes four random jumps of the

uniselector to obtain stability. At D_3, the system is again shown to be stable.

As the next example, we have two magnets y_1 and y_2 interacting. The coefficient a_{21} is set by experimenter, and the coefficient a_{12} by the uniselector. All the other coefficients are set to be zero. For each setting of a_{21}, there are then 25 different patterns of behavior due to the 25 settings of the uniselector. The results of experiment can be shown as Fig. 17.4, where the two upper curves indicate the deflections of y_1 and y_2; the lowest curve indicates the action of the uniselector. At D_1, the setting of the uniselector happens to be a stable one, and the deflections y_1 and y_2 are in the same sense. At R_1, the sign of the coefficient a_{21}

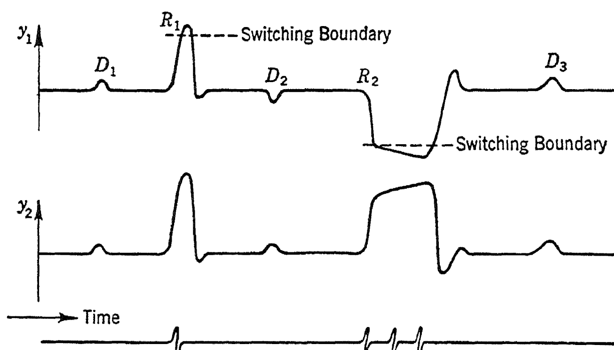


FIG. 17.4

is changed by reversing the polarity of the coil for the y_2 magnet fed with current from the y_1 magnet. This induces instability and causes the y_1 deflection to reach the switching boundary. One jump in the uniselector setting corrects the situation. At D_2, a test disturbance shows that the system is now stable, with y_1 and y_2 of opposite signs as expected. At R_2, the sign of a_{21} is again changed to that at D_1; and the uniselector setting is again unstable. Now it takes three jumps of the uniselector to reach stability. At D_3, the system is seen to be stable again, with deflections of y_1 and y_2 in the same sense.

As an example of the adaptability of an ultrastable system to even an unforseen situation, a situation not thought of until the machine has been built, we consider the course of events shown in Fig. 17.5. Here three magnets y_1, y_2, and y_3 are interacting. The behavior of the system is at first stable, as shown at D_1. The deflections y_1 and y_3 are in the same sense, but y_2 is in the opposite sense to them. Now at J, we subject the homeostat to a new, unexpected circumstance by joining the magnets y_1 and y_2 so that they must move together. This additional constraint is counter to the pattern of behavior of the prevailing setting

of the uniselector. The system becomes unstable and the consequent large deflection causes the uniselector to act. Three unstable patterns are successively rejected before the terminal stable pattern is reached, as shown at D_2. At R, the connection between the magnet y_1 and y_2 is broken; the system again becomes unstable and requires new action of the uniselector.

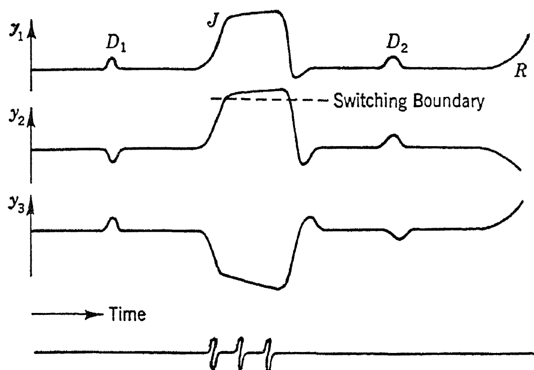


Fig. 17.5

17.3 Probability of Stability. In the previous section, we have demonstrated the characteristic adaptive behavior of the ultrastable system in seeking stable patterns of behavior under any circumstances. The question then arises: Is this searching for stability always crowned with success? What is the probability of success? If we taken an autonomous system of n variables as specified by Eq. (17.3), each value of the discrete parameter ζ gives an ordinary dynamic system. The whole range of the parameter ζ then gives an assembly of autonomous dynamic systems of n variables. We may define the probability of stability within the switching boundary in the phase space as follows. Pick a point $P(y_j)$ of the phase space, and construct an infinitesimal neighborhood dV around this point P. We can then find the differential probability dp as the fraction of the assembly of dynamical systems that will have a stable equilibrium point within the volume dV. Now integrate dp over the phase space enclosed by the switching boundary. This then gives the system's *general probability of stability* p corresponding to the specified switching boundary.

Needless to say, the actual calculation of this general probability of stability is a very difficult mathematical problem. In order to gain some ideas about this probability, Ashby made a trial calculation for an assembly of linear systems of the form

$$\frac{dy_i}{dt} = a_{ij}y_j \qquad i, j = 1, 2, \ldots, n \tag{17.5}$$

There is only one equilibrium point, the origin. The question of stability of the system is solved by considering the determinantal equation

$$|a_{ij} - \delta_{ij}\lambda| = 0 \qquad \begin{aligned} \delta_{ij} &= 1 \text{ for } i = j \\ \delta_{ij} &= 0 \text{ for } i \neq j \end{aligned} \tag{17.6}$$

If the roots λ all have negative real parts, then the system is stable. In fact, the root λ is called the latent root of the matrix a_{ij}. The probability of stability is thus the probability that the latent roots of the sample matrix a_{ij} belonging to the assembly will all have negative real parts. Ashby took the simplest possible distribution, the rectangular distribution. That is, each element a_{ij} of the matrix has an equal chance of being an integer between -9 to $+9$, inclusive. The sampling of a_{ij} was done with the aid of the table of random numbers. When $n = 1$, the probability of stability is obviously $\frac{1}{2}$. For other orders of the system, Ashby tested the stability by using Hurwitz's rule.[1] His results are shown in Table 17.1. It is seen that the probability of stability p steadily decreases with increase in the order of the system. As an approximation, the probability is $1/2^n$.

TABLE 17.1

n	Number tested	Number found stable	Per cent stable
2	320	77	24
3	100	12	12
4	100	1	1

The probability of stability can be increased by limiting the a_{ij} to favorable values. For instance, we can make all diagonal elements of the matrix have zero or negative values. Then if the variables are not interacting, the system will be always stable. The probability of stability for one variable, or $n = 1$, is now clearly unity. When $n = 2$, the probability is $\frac{3}{4}$. Ashby's test results are shown in Table 17.2.

TABLE 17.2

n	Number tested	Number found stable	Per cent stable
2	120	87	72
3	100	55	55

The probability of stability p is thus higher, but nevertheless it decreases as the number of variables is increased. From these investigations it seems reasonable to say that the probability of stability of a randomly

[1] A. Hurwitz, *Mathematischen Annalen*, **46**, 273 (1875).

constructed system steadily decreases as the system becomes more complicated. Large systems are very much more likely to be unstable than to be stable.

17.4 Terminal Fields. Following Ashby, we shall now give a definite name to the pattern of behavior at any one setting of the parameter ζ. The pattern in the phase space will be called the *field of lines of behavior.* The fields will be different for different values of the parameter. The final stable field after the switching action on the parameter will be called the *terminal field.* The action of an ultrastable system is thus mainly the search for the terminal field. It is thus of primary importance to know the average number N of switching actions necessary to reach the terminal field. This number N is very simply related to the probability of stability p of the ultrastable system. The probability of reaching the stable terminal field at the first action of the switching action is evidently the probability p itself. The probability of not reaching the terminal field is $q = 1 - p$. If the switching action is perfectly random, then the probability that the second field will be stable is still p and the probability that the second field will be unstable is q. The relative probability that the second field will be terminal is thus pq. The relative probability that the second field will not be terminal is q^2. Hence we arrive at the conclusion that the relative probability for reaching the terminal field at the mth switching action is pq^{m-1}. The average number N of switching actions for reaching the terminal field is thus

$$N = \frac{\sum_{m=1}^{\infty} mpq^{m-1}}{\sum_{m=1}^{\infty} pq^{m-1}} = \frac{(1-q)^{-2}}{(1-q)^{-1}} = \frac{1}{1-q} = \frac{1}{p} \tag{17.7}$$

When p is very small, for very large systems, the number N of switching actions necessary for reaching the stable terminal field will be very large. That is, the search for the terminal field will be long and will seem to follow a devious road.

A terminal field may be singular in the sense that only a very small fraction of the lines of behavior tend to the equilibrium point, while other lines of behavior will diverge from the equilibrium point and cross the switching boundary. Such a field will be terminal only if the line of behavior from the switching boundary happens to be one of the small former group. It will be shown presently that such singular terminal fields are not favored. If, among all possible fields of an ultrastable system, there is a certain fraction of such singular fields, the fraction of such fields used as terminal fields is much smaller. To show this, let

k be the fraction of the surface of the switching boundary from which the lines of behavior will reach the equilibrium point. For example, for the fields shown in Fig. 17.1*a* and *b*, $k = 0$. For the field shown in Fig. 17.1*c*, k is approximately $\frac{1}{2}$. For the field shown in Fig. 17.1*d*, $k = 1$. Now let $f(k)\,dk$ be the fraction of all possible fields of the ultrastable system having k between k and $k + dk$. $f(k)$ is thus the distribution function of the possible fields of the ultrastable system, and by definition

$$\int_0^1 f(k)\,dk = 1 \tag{17.8}$$

Then because only the lines of behavior that start from the part k of the switching boundary can produce a terminal field, the relative probability of having the terminal field with its k in the range from k to $k + dk$ is $kf(k)\,dk$. Thus the *distribution function* $g(k)$ of the terminal fields is related to the distribution function $f(k)$ of the possible fields of the ultrastable system by

$$g(k) = \frac{kf(k)}{\int_0^1 k'f(k')\,dk'} \tag{17.9}$$

Evidently

$$\int_0^1 g(k)\,dk = 1 \tag{17.10}$$

The terminal fields are thus definitely crowded toward larger values of k, as shown by Fig. 17.6. Singular terminal fields are thus not favored.

The actual terminal-field distribution is, however, not necessarily the potential terminal-field distribution calculated by Eq. (17.9). The reason is that any equilibrium state within a terminal field is subject to random disturbances. If the equilibrium point in the phase space is closed to the switching boundary, then even a relatively small disturbance might put the instantaneous state of the system outside the switching boundary and thus destroy that field. Therefore, for a terminal field to have a high probability of being retained under random disturbances, the equilibrium must have a central location in the switching boundary. For instance, as shown in Fig 17.7 the field C is definitely more stable than the fields A and B. The field B contains both an unstable equilibrium point and a limit cycle.

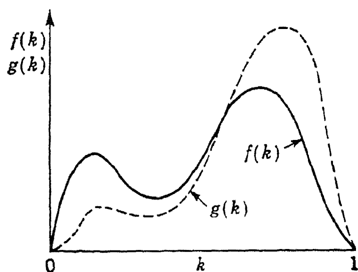


FIG. 17.6

To put this concept of stability under random disturbances into a quantitative form we have to introduce the probability σ of retaining a

terminal field after a single disturbance. If the field contains a single stable equilibrium point P, as shown in Fig. 17.7, and if the distribution function of the random disturbances from P is specified, say by a Gaussian distribution, then σ is simply the integral of this disturbance distribution function over the part of the phase space enclosed by the switching boundary. If the terminal field contains a limit cycle S, then σ is the average of the probability of retaining the field, taking as the equilibrium point each point on the limit cycle, and weighting this probability according to the proportion of time the system will spend at that

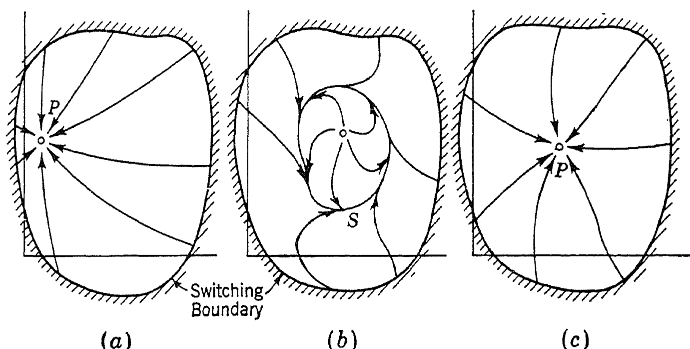


FIG. 17.7

point of the limit cycle. We can then assign a probability σ to each terminal field. Let the distribution function of the terminal fields in σ be $\varphi(\sigma)$; *i.e.*, the probability of finding a terminal field with σ in the range from σ to $\sigma + d\sigma$ is $\varphi(\sigma)\, d\sigma$. Thus

$$\int_0^1 \varphi(\sigma)\, d\sigma = 1 \tag{17.11}$$

The actual terminal-field distribution function is, however, $\psi(\sigma)$, with

$$\int_0^1 \psi(\sigma)\, d\sigma = 1 \tag{17.12}$$

To calculate $\psi(\sigma)$ in terms of $\varphi(\sigma)$, we observe that the distribution $\psi(\sigma)$, being the actual final terminal-field distribution, will not be altered by random disturbances. Secondly, after one disturbance the fields with a value of σ between σ and $\sigma + d\sigma$, $\psi(\sigma)\, d\sigma$ in relative number, will have the probability σ of being retained and the probability $(1 - \sigma)$ of being destroyed. The total relative number of fields destroyed by one random disturbance is thus

$$\int_0^1 (1 - \sigma)\psi(\sigma)\, d\sigma$$

The new terminal fields are distributed according to the potential terminal fields, *i.e.*, according to $\varphi(\sigma)$. Therefore the total relative number

of terminal fields in the range σ to $\sigma + d\sigma$ after one random disturbance is

$$\sigma\psi(\sigma)\,d\sigma + \varphi(\sigma)\,d\sigma \int_0^1 (1-\sigma')\psi(\sigma')\,d\sigma'$$

But this relative number is proportional to $\psi(\sigma)\,d\sigma$, because random disturbances should not change the final distribution $\psi(\sigma)$. Thus, if C is the constant of proportionality,

$$C\left[\sigma\psi(\sigma) + \varphi(\sigma)\int_0^1 (1-\sigma')\psi(\sigma')\,d\sigma'\right] = \psi(\sigma)$$

By integrating this expression with respect to σ from $\sigma = 0$ to $\sigma = 1$, C can be shown to be unity with Eqs. (17.11) and (17.12). Therefore we have

$$\sigma\psi(\sigma) + \varphi(\sigma)\int_0^1 (1-\sigma')\psi(\sigma')\,d\sigma' = \psi(\sigma)$$

In this equation the integral is just a constant independent of σ. Thus $\psi(\sigma)$ is proportional to $\varphi(\sigma)/(1-\sigma)$. Or, by using Eq. (17.12),

$$\psi(\sigma) = \frac{\varphi(\sigma)}{(1-\sigma)\displaystyle\int_0^1 \frac{\varphi(\sigma')\,d\sigma'}{(1-\sigma')}} \tag{17.13}$$

Equation (17.13) allows the determination of the actual terminal-field distribution from the potential terminal-field distribution. Since the potential distribution can be calculated from the distribution of all possible fields of the ultrastable system by Eq. (17.9), we can, in theory at least, obtain the actual terminal-field distribution $\psi(\sigma)$ from the specified properties of the ultrastable system. Equation (17.13) shows, furthermore, that the actual terminal-field distribution $\psi(\sigma)$ is far more crowded toward larger values of σ than $\varphi(\sigma)$ is, as expected by our previous intuitive arguments. This fact is indicated in Fig. 17.8. It should be noted that the particular type of random disturbances will affect only the calculation of the distribution function $\varphi(\sigma)$, while the relationship between $\psi(\sigma)$ and $\varphi(\sigma)$, as specified by Eq. (17.13), is independent of it and holds for any form of the distribution of disturbances.

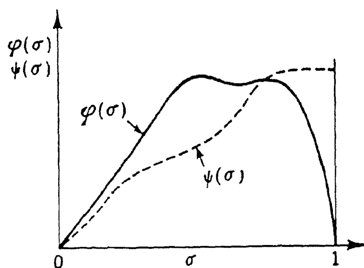


FIG. 17.8

17.5 Multistable System. As shown in the preceding section, the number N of switching actions necessary for reaching a terminal field is expected to be $1/p$, where p is the general probability of stability

of the fields of an ultrastable system. Since we have concluded that p decreases to very small values for large systems, the number N may be very large. For instance, if there are 100 variables in the system, $p \approx 1/2^{100}$, and $N = 2^{100} \approx 10^{30}$. Even if we allow for 10 switching actions per second, the average time for reaching a terminal field will still be 3×10^{19} centuries. Such a long settling time may well be considered to be infinite, and the ultrastable system will practically never reach a stable terminal field. Therefore, for large systems, or just where the principle of ultrastability for automatically reaching stable behavior is important, we find the concept to be impractical.

To remedy this situation, we must increase the probability of stability of the system. One way to do this would be a compromise. We require the design of the system to be such that the fields of the ultrastable system are limited to those which are stable under expected operating conditions. Only local and minor adjustments need be made by the switching action. In other words, the system is designed according to the conventional methods without using the principle of ultrastability. Ultrastability and switching are introduced only when trouble is expected. For instance, we can design an autopilot for an airplane according to the principles discussed in the previous chapters. But we have an apprehension that the assembly mechanic might plug the control signal from the autopilot to the aileron in a reverse manner, so that the aileron down signal actually produces aileron up motion. If the mechanic actually makes this mistake, then the autopilot will not stabilize the airplane; instead, the autopilot-airplane system will be unstable, with diverging rolling motion. However, the designer's misgiving can be quieted by introducing ultrastability just at this point. The signal connection is made to switch automatically whenever the rolling motion exceeds the specified values, the switching boundary. Then the system becomes an ultrastable system of two possible fields, one stable, one unstable. A maximum of only one switching action is necessary to reach stability, in spite of the fact that there may be a very large number of variables in the autopilot-airplane system. The point is, of course, that we do not need to leave everything to chance. We can design for stable operation under almost all conditions and greatly reduce the choice of fields of behavior by providing safeguards only for those contingencies expected to occur. Hence it is a compromise between the principles of conventional control design and the principle of ultrastability.

For living organisms, no prior knowledge of the conditions of the surroundings can be assumed, and therefore it is not possible to increase the probability of stability by limiting the choice of fields of behavior of the system. Ashby discovered a different way of increasing the probability of stability. He observed that for a very complicated system

containing a large number of variables, any single disturbance or change of operating conditions directly affects only a relatively small number of these variables. Thus if the variables directly disturbed can be isolated from the rest and form an ultrastable system by themselves, then the probability of stability for that particular type of disturbances can be greatly increased. For instance, if the number of variables directly influenced at any one disturbance is five, instead of 100 as assumed in the first paragraph of this section, the expected time of reaching its terminal field is only 3.2 seconds, with 10 switching actions per second. Thus if the 100 variables are actually divided into 20 groups of five variables each, forming 20 separate ultrastable systems, the total time required for adapting to a completely new set of operating conditions will be $20 \times 3.2 = 64$ seconds. This is a tremendous reduction from the 3×10^{19} centuries required for a completely connected ultrastable system of 100 variables.

Of course a system composed of 20 separate systems of five variables each does not have the flexibility and the rich response of a system of 100 interacting, connected variables. If, however, any one disturbance involves only five variables, then separate systems of five variables can be made to be equivalent to the system of 100 variables by making the association of the variables in any subsystem change according to the disturbance. If the disturbed variables are y_1, y_2, y_3, y_4, and y_5, then these five variables are associated to form an ultrastable system of five variables. If a later disturbance affects the variables y_2, y_5, y_{10}, y_{98}, and y_{99}, then these five variables are associated to form an ultrastable system. This phenomenon of ever-changing organization of variables into subsystems according to operating conditions is called by Ashby the *dispersion* of behavior. Dispersion can be actually achieved by making the functions $f_i(y_1,y_2, \ldots ,y_n;\zeta)$ in Eq. (17.3) zero when the point $(y_1,y_2, \ldots ,y_n)$ is within a certain region of the phase space. Then those y_i will be constants with respect to time and are tentatively only parameters with respect to other variables. In our example above, for the first disturbance, the point $(y_1,y_2, \ldots ,y_n)$ is in a region of phase space such that f_i is zero except when $i = 1, 2, 3, 4$, and 5. For the second disturbance, f_i vanishes except for $i = 2, 5, 10, 98$, and 99. It is evident that this behavior of the functions f_i only means that various thresholds exist for the derivatives dy_i/dt. Such thresholds are naturally to be expected in any real system. Therefore dispersion is something easy to obtain.

A system of ultrastable subsystems organized with the possibility of dispersion is called by Ashby a *multistable system*. A multistable system obviously has the adaptive behavior of an ultrastable system simply because it is composed of ultrastable subsystems. However it differs

from the single ultrastable system of an equal number of variables in the time required for reaching a terminal field. The multistable system has a very much shorter settling time and makes the principle of ultrastability practically realizable. Moreover, a multistable system, by making successive tentative adaptations as response to successive disturbances, shows learning by steps, or serial adaptation. This is a characteristic universally observed in living things. Furthermore, since the second and subsequent adaptations to a disturbance will certainly alter the parameter of the system, the recurrence of a disturbance identical to the first disturbance will not generally reproduce the behavior of the system at the first adaptation. This is indeed the dispersion of behavior. In other words, as the system "grows older" it becomes "wiser," and its behavior is appropriate for a wider range of activities than just for coping with the single prevailing disturbance.

CHAPTER 18

CONTROL OF ERROR

In the preceding chapter, we have shown how the principle of ultrastability can make the control system insensitive to accidental errors and occasional failures of the components by the simple device of changing the characteristics of the system whenever instability occurs. Since an ultrastable system will automatically seek stability, the control system, when designed, actually embodies unstable fields of behavior as well as stable fields of behavior. In other words, during the design of an ultrastable system, we make no attempt to distinguish stability from instability, to separate the right fields of behavior from the wrong fields of behavior. Errors of behavior are merely treated as a probability, but otherwise unspecified. In this chapter, we shall approach the reliability of a complex control system from a different point of view: we shall specifically introduce errors into the system and ask how the system should be designed so that the system will give satisfactory performance in spite of the errors. That is, we wish to know how to control the error.

This subject of control of error is now in its early period of development. The control of error can be discussed for only the most elementary operations and the present theory is wholly due to J. von Neumann.[1] Our discussion in this chapter is then an exposition of Neumann's work. Its purpose is to serve as an introduction to this very important topic and to indicate the need for much further investigation.

18.1 Reliability by Duplication. It is common knowledge that the reliability of a system can generally be increased by the simple expedient of duplication. For instance, if a simple system as shown in Fig. 18.1*a* has the characteristic that when it fails to operate, it merely gives no output, then to guard against probability of failure, we can duplicate the system with identical systems and put all in parallel, as shown in Fig. 18.1*b*. If the probability of failure of the original system is p, a number between 0 and 1, then the probability of failure of each individual unit in the parallel system is also p. If the units in the parallel system are independent of each other, the failure of the parallel system will

[1] J. von Neumann, "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components," Printed notes of lectures given at the California Institute of Technology, Pasadena, Calif., 1952.

occur only when every unit fails. The probability of failure of the parallel system is thus p^n. By increasing the number n of duplications, we can make this probability very small indeed.

However, failure of a component of a control system generally does not cause zero output. Instead, the effect is much more damaging: the system will still give an output, but the wrong output. Then simple duplication of the system, as described above, will not be effective, because a wrong output when mixed with a correct output results in a wrong output. Hence, in such a malfunction of the system, as opposed to failure, the probability of the malfunction of the parallel system

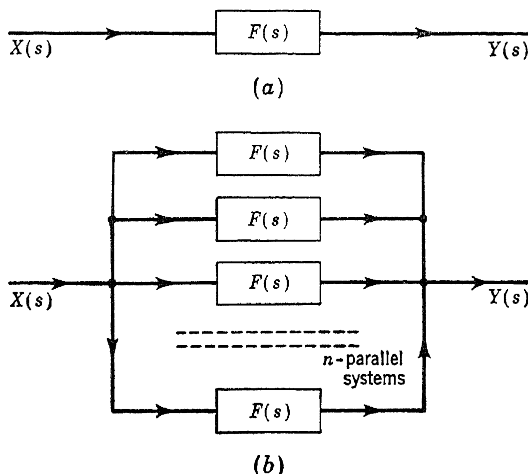


FIG. 18.1

is equal to the probability of the malfunction of the single system; and no improvement of reliability results. Therefore the problem of control of error is a deeper and more difficult problem than it seems to be at first sight. Nevertheless, as will be seen later, the principle of duplication, *i.e.*, the necessity of increasing the number of components, remains basic. What has to be discovered is a new organization of these components for controlling the error, because the simple parallel organization of Fig. 18.1*b* is not effective.

18.2 Basic Elements. To simplify the analysis, we shall not discuss the case where the input and the output are continuous functions, but we choose an elementary component with input and output capable of only two discrete values, 0 or 1, *i.e.*, the input is either on (activated) or off, the output is either on (activated) or off. The characteristic of the element is then determined by the relation between the state of the input and the state of the output. We shall always assume a single output, but there may be more than one input to the element. There

is also a time lag between the output and the inputs in the sense that the output is activated only at a certain specified time interval after the inputs are activated. Such an element then has the properties of a relay circuit.

To describe the characteristic of the element, we introduce four kinds of input: excitory input; inhibitory input; permanently excitory, or live, input; and a kind of input which is never activated, or a grounded input. They are represented by symbols according to Fig. 18.2. The element is then represented by a circle with inputs on the left side and the single output on the right. Within the circle there is a number k, meaning that the number of activated excitory inputs must be greater than the number of activated inhibitory inputs by k or more, for the output to be on. Thus Fig. 18.3*a* represents an element whose output is on only if both inputs a and b are on. It can be called an ab element. Figure 18.3*b* represents an element whose output is on if either a or b or both are on. It can be called an $a + b$ element. Figure 18.3*c* represents an element whose output is on only if the input a is off. It can

Excitory Input

Inhibitory Input

Live Input

Grounded Input

Fig. 18.2

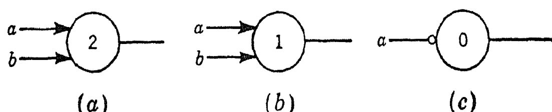


Fig. 18.3

be called an a^{-1} element. Incidentally, if we consider the inputs to be the conditions for a given statement to be true (on) or false (off), then the three elements of Fig. 18.3 represent the three fundamental operations of Boolean algebra. In a digital computer using binary numbers, these elements are the basic elements for its operation. If memory is necessary for the computing operation, such a device can be made from the second element of Fig. 18.3 by feedback as shown in Fig. 18.4. Once the input a is activated, the output will be on, even if a is later turned off.

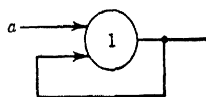
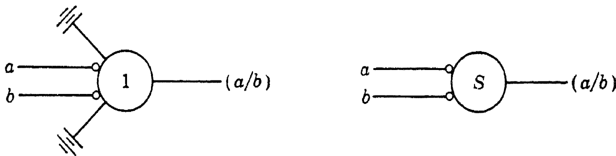


Fig. 18.4

For later discussions, having three basic elements to deal with is a disadvantage. However, all three are really special cases of one fundamental element. Consider the element represented by Fig. 18.5, called the *Scheffer stroke*. Since the two live inputs are always excited, we can eliminate them from the diagram and represent the element as shown in the right diagram. The output of the Scheffer stroke is thus on if neither a nor b is on, or if either a or b is on; but the output is off if both

a and b are on. The three basic elements of the last paragraph can be constructed out of the Scheffer stroke as shown in Fig. 18.6. Of course both the ab element and the $a + b$ element involve two Scheffer-stroke elements in series; thus the time lag will be twice as large as for the a^{-1} element having only one Scheffer stroke. But since our notion of time lag merely indicates that inputs precede output, the exact magnitude of time lag is of no consequence. Operations differing only in time lag will be considered equivalent. Therefore all three basic operations can be represented by a single Scheffer-stroke element.



Scheffer Stroke
Fig. 18.5

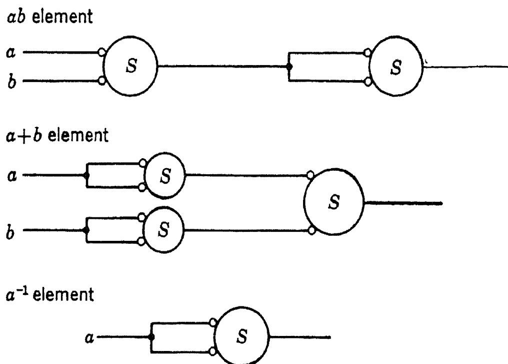


Fig. 18.6

The choice of a single fundamental element is, however, not unique. There can be other choices besides the Scheffer stroke. But our choice is convenient for the following discussions. We shall consider first the control of error in an operation represented by the Scheffer stroke; then any complex operation built up from Scheffer strokes can be similarly designed.

18.3. Method of Multiplexing. Applying the concept of duplication for improving the reliability, we substitute any single input by a *bundle of inputs* with n individual lines. Thus in the system of the single Scheffer stroke, we would have n lines for the a input, labelled a_i, where $i = 1, 2, \ldots, n$; n lines for the b input, labelled b_i for $i = 1, 2, \ldots, n$; and n lines in the *output bundle*. We then specify the fraction δ, where $0 < \delta < \frac{1}{2}$, such that if $(1 - \delta)n$ lines of the output bundle are on or off,

the output as a whole is considered to be on or off. If δn lines of the output bundle are on or off, then the output as a whole is considered to be off or on, respectively. Any intermediate value is considered to be a malfunction. δ is thus the *fiduciary level.* The problem is how to construct the system, using Scheffer-stroke elements, such that the probability of malfunction can be reduced with a specified probability of errors in the input bundles and a specified probability of malfunction of the individual Scheffer-stroke elements.

A first approach to the problem would be to take a line a_i from the a input bundle and a line b_i from the b input bundle, and to use these as

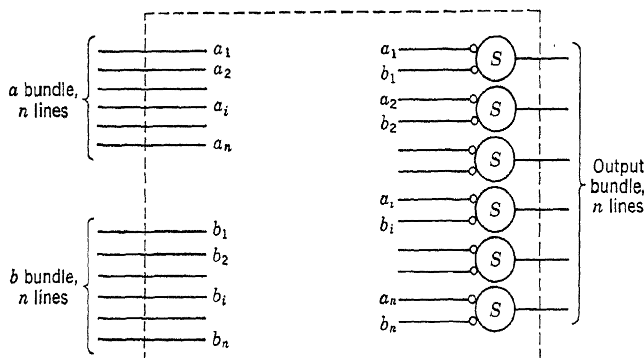


FIG. 18.7

the two inputs to a Scheffer-stroke element. The organization of the system is thus as shown in Fig. 18.7. It is evident that if almost all lines of both input bundles are on, then almost all lines of the output bundle will be off. If almost all lines of both input bundles are off, then almost all lines of the output bundle will be on. This over-all behavior seems to be satisfactory. However, a more careful consideration will show that it is not so. Since, for output of Scheffer stroke to be off, both inputs have to be on, an error either in the a bundle or in the b bundle will be sufficient to cause an error in the output bundle. Therefore, if the output is supposed to be off, then the error in output is the *sum* of errors in the input bundles. Similarly, if the output is supposed to be on and only one input bundle is off, then the error in output is the same as that in the input bundle. If the output is supposed to be on and both input bundles are off, then an error in the output bundle requires a *simultaneous* error in both input bundles; and therefore the error level in the output bundle is less than the error level in the input bundles. Therefore the error level in the inputs is not maintained at the output. Some situations result in a magnification of error, while others result in a reduction of error. There is then a dispersion of error. This is undesir-

able, because dispersion of error causes the number of activated lines of the output bundle to drift to the undetermined region between δn lines and $(1 - \delta)n$ lines, and thus increases the chance of malfunction.

To suppress the dispersion of error, we can introduce a *restoring component* of the system as follows. We take each line of the output bundle from the *executive component* of Fig. 18.7, executive in the sense of carrying out the Scheffer-stroke operation for the over-all system, and split it into two lines. We then obtain $2n$ lines. Then we permutate these $2n$ lines so that the order of lines is now made random. Taking the successive pairs of lines and using them as inputs to a Scheffer-stroke element, we then again obtain a bundle of n output lines. This is

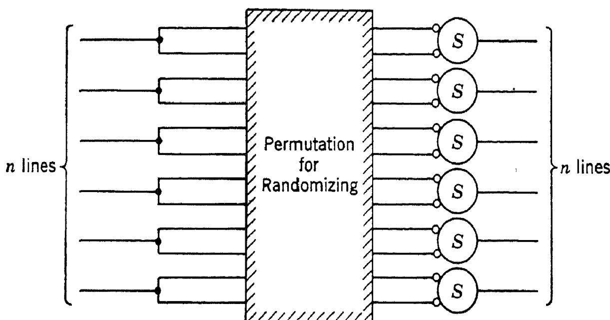


Fig. 18.8

indicated diagrammatically in Fig. 18.8. Let there be $\alpha_0 n$ activated lines in the original bundle. The fraction of the output lines not activated is clearly $\alpha_0\alpha_0 = \alpha_0^2$. The fraction α_1 of the output lines activated is thus

$$\alpha_1 = 1 - \alpha_0^2 \tag{18.1}$$

If α_0 is the probability of having the original lines activated, then, provided n is large, the probability of having the transformed lines activated is α_1. This is not a restoring component yet. But if we put two such units in series, the probability α_2 of having the final lines activated is

$$\alpha_2 = 1 - \alpha_1^2 = 1 - (1 - \alpha_0^2)^2 = 2\alpha_0^2 - \alpha_0^4 \tag{18.2}$$

The series system is now a restoring component for the following reason. Fig. 18.9 shows the relationship between α_2 and α_0. α_2 is equal to α_0 when

$$\alpha_0^4 - 2\alpha_0^2 + \alpha_0 = 0$$

or when $\alpha_0 = 0, \frac{1}{2}(\sqrt{5} - 1)$, or 1. Thus if α_0 lies between 0 and

$$\tfrac{1}{2}(\sqrt{5} - 1) = 0.618034$$

α_2 is smaller than α_0; if α_0 lies between $\frac{1}{2}(\sqrt{5} - 1)$ and 1, α_2 is greater than α_0. Therefore the action of the restoring component is to drive the probability of activating the output towards the limits 0 and 1, and thus to reduce the dispersion of error caused by the executive component.

On the basis of the preceding discussion, then, our system for error control consists of an executive component of n individual Scheffer-stroke elements, followed by a restoring component composed of two units of Fig. 18.8 in series, each made of n Scheffer-stroke elements and a randomizing device. Therefore, for each Scheffer-stroke element of perfect accuracy, we have to expand the system to a complex system of $3n$ Scheffer-stroke elements. With any fixed fiduciary level δ and with given probabilities of error in the input lines and the Scheffer-stroke elements, we shall see that we can make the reliability of the over-all system as high as we please by increasing n. Basically then, the principle is still reliability by duplication. However, our analysis in this section gives us a specific plan of organizing these elements. This particular method of synthesizing a reliable system out of unreliable elements is called the method of multiplexing by von Neumann.

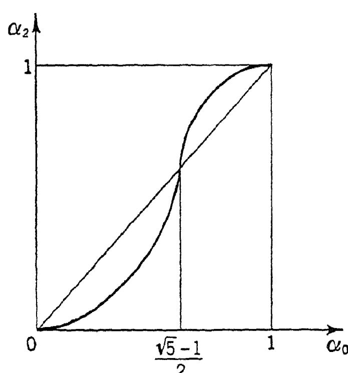


Fig. 18.9

18.4 Error in Executive Component. We shall now compute the error of the multiplexed Scheffer-stroke system described in the preceding section and show that the error is indeed controlled. We observe first that the direct source of error is the individual Scheffer-stroke element itself, in either the executive component or in the restoring component. Let the probability for each individual element of making a mistake be ϵ. If we assume that there are r parallel Scheffer-stroke elements and that they are independent in their operation, then the probability for the elements to fail when in the system is still ϵ. Therefore, of the r parallel elements, on the average ϵr will make mistakes. This number is also the most likely number of mistakes in the n parallel elements. The probability of making some other number of mistakes is less. In fact, the problem of determining the probability $p_0(\rho,\epsilon,r)$ of having ρ mistakes in r parallel elements with an individual probability of failure equal to ϵ is the classical problem of random sampling. For large r, it is known[1] that

[1] See for instance H. Margenau and G. M. Murphy, "The Mathematics of Physics and Chemistry," p. 422, D. Van Nostrand Company, Inc., New York, 1943.

$$p_0(\rho,\epsilon,r) \approx \frac{1}{\sqrt{2\pi}\sqrt{\epsilon(1-\epsilon)r}} e^{-\frac{1}{2}\frac{(\rho-\epsilon r)^2}{\epsilon(1-\epsilon)r}} \tag{18.3}$$

Therefore the probability distribution $p_0(\rho,\epsilon,r)$ is a normal distribution with a mean at ϵr and a mean deviation equal to $\sqrt{\epsilon(1-\epsilon)r}$.

The other source of error in the executive component is the misfitting of lines from the input bundles into the individual Scheffer-stroke element. For instance, if a fraction ξ of the a input bundle of Fig. 18.7 is activated and if a fraction η of the b input bundle is activated, we would expect a similar fraction of the output bundle to be inhibited. But if the particular a_i for ith element is activated and b_i happens to be nonactivated, the output of the ith element is still activated, even with no mistake on the part of the element itself. Let ζ be the fraction of the output bundle of the executive component that is activated. Then the number of output lines effectively inhibited is $(1-\zeta)n$. The total number of activated lines in the a input bundle is ξn, in the b input bundle it is ηn. The number of effective, or properly fitted, a-input lines is, however, only $(1-\zeta)n$. The difference $[\xi-(1-\zeta)]n$ is ineffective. The number of ineffective b-input lines is $[\eta-(1-\zeta)]n$. Therefore the number of effective output lines is $(1-\zeta)n$, the number of ineffective output lines due to misfitting activated a input is $[\xi-(1-\zeta)]n$, the number of ineffective output lines due to misfitting activated b input is $[\eta-(1-\zeta)]n$, and, finally, the number of ineffective output lines due to nonactivated input lines is the remainder

$$\{1-(1-\zeta)-[\xi-(1-\zeta)]-[\eta-(1-\zeta)]\}n = (2-\xi-\eta-\zeta)n$$

The number of possible effective combinations of such a classification of output is thus[1]

$$\frac{n!}{[(1-\zeta)n]!\{[\xi-(1-\zeta)]n\}!\{[\eta-(1-\zeta)]n\}![(2-\xi-\eta-\zeta)n]!}$$

On the input side, the number of possible combinations of n a-input lines, with ξn activated, $(1-\xi)n$ nonactivated, is

$$\frac{n!}{(\xi n)![(1-\xi)n]!}$$

The number of possible combinations of n b-input lines, with ηn activated, $(1-\eta)n$ nonactivated, is

$$\frac{n!}{(\eta n)![(1-\eta)n]!}$$

Therefore the probability p_1 of having a fraction ζ of the output activated with fractions ξ and η of the inputs activated and with perfect

[1] See for instance H. Margenau and G. M. Murphy, *op. cit.*, p. 415.

operation of the individual Scheffer-stroke elements, is

$$p_1(\xi,\eta,\zeta;n) = \frac{\dfrac{n!}{[(1-\zeta)n]!\{[\xi-(1-\zeta)]n\}!\{[\eta-(1-\zeta)]n\}![(2-\xi-\eta-\zeta)n]!}}{\dfrac{n!}{[\xi n]![(1-\xi)n]!}\,\dfrac{n!}{[\eta n]![(1-\eta)n]!}}$$
$$= \frac{[\xi n]![(1-\xi)n]![\eta n]![(1-\eta)n]!}{[(1-\zeta)n]!\{[\xi-(1-\zeta)]n\}!\{[\eta-(1-\zeta)]n\}![(2-\xi-\eta-\zeta)n]!n!} \quad (18.4)$$

Clearly, for the enumeration to have meaning, none of the four classes of output lines discussed in the preceding paragraph can have less than zero number. Whenever that happens, the probability drops to zero. That is, p_1 is zero whenever the following conditions are violated,

$$\left.\begin{aligned} 1-\zeta &> 0 \\ \xi-(1-\zeta) &> 0 \\ \eta-(1-\zeta) &> 0 \\ \text{and} \qquad 2-\xi-\eta-\zeta &> 0 \end{aligned}\right\} \quad (18.5)$$

We shall now simplify the expression of Eq. (18.4) under the assumption that n is large. When n is large, the factorials can be approximated by their asymptotic value, given by Stirling's formula

$$n! \approx \sqrt{2\pi}\, e^{-n} n^{n+\frac{1}{2}} \quad (18.6)$$

By using Eq. (18.6) we can write the approximate expression for p_1 as

$$p_1(\xi,\eta,\zeta;n) = \frac{1}{\sqrt{2\pi n}}\sqrt{a}\, e^{-\theta n} \quad (18.7)$$

where

$$a = \frac{\xi(1-\xi)\eta(1-\eta)}{(\xi+\zeta-1)(\eta+\zeta-1)(1-\zeta)(2-\xi-\eta-\zeta)} \quad (18.8)$$

and

$$\begin{aligned}\theta = {} & (\xi+\zeta-1)\log(\xi+\zeta-1) + (\eta+\zeta-1)\log(\eta+\zeta-1) \\ & + (1-\zeta)\log(1-\zeta) + (2-\xi-\eta-\zeta)\log(2-\xi-\eta-\zeta) \\ & - \xi\log\xi - (1-\xi)\log(1-\xi) - \eta\log\eta - (1-\eta)\log(1-\eta) \end{aligned} \quad (18.9)$$

Differentiating θ with respect to ζ, we have

$$\frac{\partial\theta}{\partial\zeta} = \log\frac{(\xi+\zeta-1)(\eta+\zeta-1)}{(1-\zeta)(2-\xi-\eta-\zeta)} \quad (18.10)$$

and

$$\frac{\partial^2\theta}{\partial\zeta^2} = \frac{1}{\xi+\zeta-1} + \frac{1}{\eta+\zeta-1} + \frac{1}{1-\zeta} + \frac{1}{2-\xi-\eta-\zeta} \quad (18.11)$$

We find by using these equations that at $\zeta = 1 - \xi\eta$, $\theta = \frac{\partial\theta}{\partial\zeta} = 0$. Furthermore, because of the conditions of Eq. (18.5), $\frac{\partial^2\theta}{\partial\zeta^2}$ is always positive. Therefore the only zero of θ is at $\zeta = 1 - \xi\eta$. Then if n is very large, the negative exponential in Eq. (18.7) shows that we need only consider θ near its zero. But at the zero of θ, where $\zeta = 1 - \xi\eta$,

$$\frac{\partial^2\theta}{\partial\zeta^2} = \frac{1}{\xi(1-\eta)} + \frac{1}{\eta(1-\xi)} + \frac{1}{\xi\eta} + \frac{1}{(1-\xi)(1-\eta)} = \frac{1}{\xi(1-\xi)\eta(1-\eta)}$$

Thus near $\zeta = 1 - \xi\eta$, θ is approximated by

$$\theta \sim \frac{1}{2}\frac{[\zeta - (1-\xi\eta)]^2}{\xi(1-\xi)\eta(1-\eta)} \tag{18.12}$$

When n is large, a is a relatively slowly varying function of ζ in comparison to the exponential of Eq. (18.7). Thus we can take the value of a at the point $\zeta = 1 - \xi\eta$. Or

$$a \sim \frac{1}{\xi(1-\xi)\eta(1-\eta)} \tag{18.13}$$

Therefore, finally, the approximate expression for $p_1(\xi,\eta,\zeta;n)$ with very large n is

$$p_1(\xi,\eta,\zeta;n) \approx \frac{1}{\sqrt{2\pi\xi(1-\xi)\eta(1-\eta)n}}\, e^{-\frac{1}{2}\frac{[\zeta-(1-\xi\eta)]^2 n}{\xi(1-\xi)\eta(1-\eta)}} \tag{18.14}$$

Hence p_1 is also a normal distribution with respect to ζ. The mean is at $(1 - \xi\eta)n$, and the mean deviation is $\sqrt{\xi(1-\xi)\eta(1-\eta)n}$.

We shall make a final modification of the probability $p_1(\xi,\eta,\zeta;n)$ by passing to the continuous distribution function $W(\zeta;\xi,\eta;n)$ for large values of n. If $W(\zeta;\xi,\eta:n)\,d\zeta$ is the probability of having the number of activated output lines from ζn to $\zeta n + 1 = n(\zeta + 1/n)$, then $d\zeta = 1/n$, and this probability is exactly $p_1(\xi,\eta,\zeta;n)$. Therefore

$$W(\zeta;\xi,\eta;n) = np_1 = \frac{1}{\sqrt{2\pi\xi(1-\xi)\eta(1-\eta)/n}}\, e^{-\frac{1}{2}\left[\frac{\zeta-(1-\xi\eta)}{\sqrt{\xi(1-\xi)\eta(1-\eta)/n}}\right]^2} \tag{18.15}$$

$W(\zeta;\xi,\eta;n)\,d\zeta$ in general is then the probability of having the fraction of activated output lines between ζ and $\zeta + d\zeta$, with specified fractions ξ and η of the input bundles activated. The size of bundles is determined by the number of lines n. W is a Gaussian distribution with the mean $1 - \xi\eta$ and the mean derivation $\sqrt{\xi(1-\xi)\eta(1-\eta)/n}$. An equivalent way of expressing the result is to write

$$\zeta = (1 - \xi\eta) + \sqrt{\frac{\xi(1-\xi)\eta(1-\eta)}{n}}\, y \tag{18.16}$$

where y denotes a random variable distributed according to the normal Gaussian distribution, with the mean zero and the mean deviation equal to unity. Then Eq. (18.16) states that ζ is distributed according to the Gaussian distribution with the mean $1 - \xi\eta$ and the mean deviation $\sqrt{\xi(1-\xi)\eta(1-\eta)/n}$. Therefore Eqs. (18.15) and (18.16) express the same fact. Equation (18.16) is, however, often more convenient to use.

We can now combine the two sources of errors and add to the ζ distribution of Eq. (18.16) the effects of imperfect individual Scheffer-stroke elements. We can rewrite Eq. (18.3) in a way similar to Eq. (18.16):

$$\rho = \epsilon r + \sqrt{\epsilon(1-\epsilon)r}\, y \tag{18.17}$$

Now in our executive component there are two classes of the Scheffer-stroke elements. One type, $\zeta\eta$ in number, is supposed to have activated output. A mistake will decrease the number of activated lines by one. The q mistakes in this type of elements are given by Eq. (18.17) as

$$q = \epsilon\zeta n + \sqrt{\epsilon(1-\epsilon)\zeta n}\, y \tag{18.18}$$

The other type, $(1 - \zeta)n$ in number, is supposed to be nonactivated. A mistake will increase the number of activated lines by one. The q' mistakes in this type of elements are distributed as

$$q' = \epsilon(1-\zeta)n + \sqrt{\epsilon(1-\epsilon)(1-\zeta)n}\, y \tag{18.19}$$

$q' - q$ is then the additional number of activated output lines due to errors of the elements themselves. According to Eqs. (18.18) and (18.19), we have

$$q' - q = 2\epsilon(\tfrac{1}{2} - \zeta)n + \sqrt{\epsilon(1-\epsilon)(1-\zeta)n}\, y - \sqrt{\epsilon(1-\epsilon)\zeta n}\, y \tag{18.20}$$

The last two terms of the above equation are the difference of two normally distributed random variables. We shall see presently that the difference is again a normally distributed random variable.

Consider two normally distributed random variables z_1 and z_2 with the mean zero, and the mean deviations σ_1 and σ_2, respectively. Then

$$\left.\begin{aligned} z_1 &= \sigma_1 y \\ z_2 &= \sigma_2 y \end{aligned}\right\} \tag{18.21}$$

Or, with $W_1(z_1)$ denoting the probability distribution function of z_1, $W_2(z_2)$ the probability distribution function of z_2,

$$W_1(z_1) = \frac{1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{z_1}{\sigma_1}\right)^2}$$

$$W_2(z_2) = \frac{1}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{z_2}{\sigma_2}\right)^2}$$

Now if the two random variables are *independent* of each other, the joint probability of having z_1 in the range z_1 to $z_1 + dz_1$ and z_2 in the range z_2 to $z_2 + dz_2$ is

$$W_1(z_1) W_2(z_2)\, dz_1\, dz_2$$

We now introduce new variables x_1 and x_2 defined by

$$\begin{aligned} x_1 &= z_1 - z_2 \\ x_2 &= z_1 + z_2 \\ z_1 &= \tfrac{1}{2}(x_1 + x_2) \\ z_2 &= \tfrac{1}{2}(x_2 - x_1) \end{aligned}$$

The joint probability in the new variables is then

$$\tfrac{1}{2} W_1 \left(\frac{x_1 + x_2}{2} \right) W_2 \left(\frac{x_2 - x_1}{2} \right) dx_1\, dx_2$$

By integrating this joint probability with respect to x_2 from $-\infty$ to ∞, we obtain the probability $W(x_1)\, dx_1$, where $W(x_1)$ is the probability distribution of $x_1 = z_1 - z_2$. Thus

$$\begin{aligned} W(x_1) &= \tfrac{1}{2} \int_{-\infty}^{\infty} W_1 \left(\frac{x_1 + x_2}{2} \right) W_2 \left(\frac{x_2 - x_1}{2} \right) dx_2 \\ &= \frac{1}{4\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\left(\frac{z_1 + x_2}{2\sigma_1} \right)^2 + \left(\frac{x_2 - x_1}{2\sigma_2} \right)^2 \right]} dx_2 \\ &= \frac{1}{4\pi\sigma_1\sigma_2} e^{-\frac{1}{2} \frac{x_1^2}{\sigma_1^2 + \sigma_2^2}} \int_{-\infty}^{\infty} e^{-\xi^2} \frac{d\xi}{\sqrt{(1/8\sigma_1^2) + (1/8\sigma_2^2)}} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}} e^{-\frac{1}{2}\left(\frac{x_1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \right)^2} \end{aligned}$$

Therefore we can write

$$z_1 - z_2 = \sqrt{\sigma_1^2 + \sigma_2^2}\, y \tag{18.22}$$

By integrating the joint probability with respect to x_1, we can show that

$$z_1 + z_2 = \sqrt{\sigma_1^2 + \sigma_2^2}\, y \tag{18.23}$$

Hence the difference or the sum of two normally distributed and independent random variables is again a normally distributed random variable with the square of the mean deviation equal to the sum of the squares of the mean deviations of the original random variables. This ability to add or subtract normally distributed random variables with zero mean values is to be expected, since they have equal probabilities for positive or negative values.

With the relation of Eq. (18.22), we can write the result of Eq. (18.20) as

$$q' - q = 2\epsilon(\tfrac{1}{2} - \zeta)n + \sqrt{\epsilon(1 - \epsilon)n}\, y$$

Let $(q' - q)/n = \Delta\zeta$ be the correction of the fraction ζ for imperfect Scheffer-stroke elements; then we have

$$\Delta\zeta = 2\epsilon(\tfrac{1}{2} - \zeta) + \sqrt{\frac{\epsilon(1-\epsilon)}{n}}\, y \tag{18.24}$$

We can now combine Eqs. (18.16) and (18.24), and the corrected fraction ζ' of activated output lines is

$$\begin{aligned}\zeta' &= \zeta + \Delta\zeta = \zeta + 2\epsilon(\tfrac{1}{2} - \zeta) + \sqrt{\frac{\epsilon(1-\epsilon)}{n}}\, y \\ &= (1 - \xi\eta) + 2\epsilon(\xi\eta - \tfrac{1}{2}) + (1 - 2\epsilon)\sqrt{\frac{\xi(1-\xi)\eta(1-\eta)}{n}}\, y \\ &\qquad + \sqrt{\frac{\epsilon(1-\epsilon)}{n}}\, y\end{aligned} \tag{18.25}$$

The last two terms of Eq. (18.25) represent the sum of two independent, normally distributed random variables. We can thus use Eq. (18.23). Therefore, finally, writing ζ in place of ζ', we have from Eq. (18.24)

$$\begin{aligned}\zeta &= (1 - \xi\eta) + 2\epsilon(\xi\eta - \tfrac{1}{2}) \\ &\qquad + \sqrt{\frac{(1-2\epsilon)^2\xi(1-\xi)\eta(1-\eta) + \epsilon(1-\epsilon)}{n}}\, y\end{aligned} \tag{18.26}$$

where y is a normally distributed random variable with the mean equal to zero and the mean deviation equal to unity. Equation (18.26) specifies the performance of the executive component of our multiplexed Scheffer-stroke system with the fractions ξ, η, and ζ of the inputs and output activated and with ϵ as the probability of failure of the individual Scheffer-stroke elements.

18.5 Error of Multiplexed Systems. After having calculated the performance of the executive component of our multiplexed Scheffer-stroke system, we find that the rest of the computation is very easy. Each unit (Fig. 18.8) of the restoring component is really equivalent to the executive component. For the first stage of the restoring component, the inputs are the split output lines from the executive component. Thus instead of two different fractions ξ and η, we have the same fraction ζ. Therefore, if μ is the fraction of activated output of the first stage, then, according to Eq. (18.26),

$$\mu = (1 - \zeta^2) + 2\epsilon(\zeta^2 - \tfrac{1}{2}) + \sqrt{\frac{(1-2\epsilon)^2\zeta^2(1-\zeta)^2 + \epsilon(1-\epsilon)}{n}}\, y \tag{18.27}$$

Similarly if ν is the fraction of activated output of the second stage of the restoring component, then

$$\nu = (1 - \mu^2) + 2\epsilon\,(\mu^2 - \tfrac{1}{2}) + \sqrt{\frac{(1-2\epsilon)^2\mu^2(1-\mu)^2 + \epsilon(1-\epsilon)}{n}}\, y \tag{18.28}$$

Equations (18.27) and (18.28) have a first term identical with Eq. (18.1). The additional terms come from the imperfect elements and from the statistical distribution of errors.

With any specified ξ, η, ϵ, and n, Eqs. (18.26) to (18.28) enable us to compute the distribution function of ν, the fraction of activated output lines of the complete Scheffer-stroke system. We can make this somewhat clearer by reverting to the notation of probability distribution functions. Thus, for instance, Eq. (18.26) is equivalent to

$$W(\zeta;\xi,\eta;n) = \frac{\exp\left\{-\frac{1}{2}\left[\frac{\zeta - [(1-\xi\eta) + 2\epsilon(\xi\eta - \frac{1}{2})]}{\sqrt{\frac{(1-2\epsilon)^2\xi(1-\xi)\eta(1-\eta) + \epsilon(1-\epsilon)}{n}}}\right]^2\right\}}{\sqrt{2\pi\frac{(1-2\epsilon)^2\xi(1-\xi)\eta(1-\eta) + \epsilon(1-\epsilon)}{n}}}$$

The probability distribution function of ν, $W(\nu;\xi,\eta;n)$, is thus the result of integrating with respect to ζ and μ the joint probability of ζ, μ, and ν. Thus

$$W(\nu;\xi,\eta;n) = \frac{1}{(2\pi)^{\frac{3}{2}}}\frac{1}{\sqrt{\frac{(1-2\epsilon)^2\xi(1-\xi)\eta(1-\eta)+\epsilon(1-\epsilon)}{n}}}\int_{-\infty}^{\infty} d\mu$$

$$\int_{-\infty}^{\infty}\frac{d\zeta}{\sqrt{\frac{(1-2\epsilon)^2\zeta^2(1-\zeta)^2+\epsilon(1-\epsilon)}{n}\,\frac{(1-2\epsilon)^2\mu^2(1-\mu)^2+\epsilon(1-\epsilon)}{n}}}$$

$$\exp\left\{-\frac{1}{2}\left[\frac{\zeta - \{(1-\xi\eta)+2\epsilon(\xi\eta-\frac{1}{2})\}}{\sqrt{\frac{(1-2\epsilon)^2\xi(1-\xi)\eta(1-\eta)+\epsilon(1-\epsilon)}{n}}}\right]^2 - \frac{1}{2}\left[\frac{\mu - \{(1-\zeta^2)+2\epsilon(\zeta^2-\frac{1}{2})\}}{\sqrt{\frac{(1-2\epsilon)^2\zeta^2(1-\zeta)^2+\epsilon(1-\epsilon)}{n}}}\right]^2 - \frac{1}{2}\left[\frac{\nu - \{(1-\mu^2)+2\epsilon(\mu^2-\frac{1}{2})\}}{\sqrt{\frac{(1-2\epsilon)^2\mu^2(1-\mu)^2+\epsilon(1-\epsilon)}{n}}}\right]^2\right\} \quad (18.29)$$

We shall now show that under proper conditions we can obtain almost perfect performance of the multiplexed Scheffer-stroke system by increasing n. Consider a given fiduciary level δ. Perfect performance requires that $\nu \leq \delta$ for the nonactivation of output is implied by $\xi \geq 1 - \delta$ and $\eta \geq 1 - \delta$ for the activation of both inputs; and that $\nu \geq 1 - \delta$ by either $\xi \leq \delta$ and $\eta \geq 1 - \delta$ or $\xi \geq 1 - \delta$ and $\eta \leq \delta$. Let us assume that n is so large and ϵ so small that terms of order ϵ and $1/\sqrt{n}$ can be neglected in Eqs. (18.26) to (18.28). Then

$$\zeta \approx (1 - \xi\eta) \qquad \mu \approx 1 - \zeta^2 \qquad \nu \approx 1 - \mu^2$$

Or

$$\nu \approx 1 - (2\xi\eta - \xi^2\eta^2)^2 \qquad \text{when } n \gg 1, \ \epsilon \ll 1 \tag{18.30}$$

Now let $\xi = 1 - \alpha$, $\eta = 1 - \beta$, and $\alpha, \beta \leq \delta$; so that $\xi \geq 1 - \delta$ and $\eta \geq 1 - \delta$. Then Eq. (18.30) gives

$$\nu \approx 2(\alpha^2 + \beta^2) + \cdots$$

Hence $\nu = 0(\delta^2)$. Similarly, Eq. (18.30) gives $\nu = 1 - 0(\delta^2)$ for $\xi \leq \delta$ and $\eta \geq 1 - \delta$, or $\xi \geq 1 - \delta$ and $\eta \leq \delta$. Furthermore, Eq. (18.30) also gives $\nu = 1 - 0(\delta^4)$ for $\xi \leq \delta$ and $\eta \leq \delta$. Therefore perfect reliability of the multiplexed Scheffer-stroke system can be indeed obtained with $n \to \infty$, provided ϵ and δ are small.

When n is large but not infinite, the calculation is somewhat tedious because of the necessity of evaluating the integral of Eq. (18.29). Although the asymptotic values of the integral can be determined by the classical methods, we shall not enter into this calculation here. Instead we shall cite an example from von Neumann where $\delta = 0.07$, *i.e.*, activation of at least 93 per cent of the lines of a bundle represents a positive message; activation of at most 7 per cent of the lines represents a negative message. He then found that the probability ϵ of malfunction of the individual Scheffer-stroke elements must be less than 0.0107 for controlling the error. For $\epsilon \geq 0.0107$, the probability of malfunction of the over-all system cannot be made arbitrarily small by increasing n. For $\epsilon = 0.005$, or $\frac{1}{2}$ per cent chance of failure, von Neumann gave the numerical results in Table 18.1. It is seen that for as many as 1000 lines in a bundle, the reliability is rather poor. In fact it is inferior to the original 1 per cent of ϵ. But a 25-fold increase in n will give extreme reliability.

For systems organized originally in Scheffer-stroke elements, the technique of multiplexing discussed in the preceding section can also be applied without change. We replace each Scheffer-stroke element in

TABLE 18.1

$\delta = 0.07$, $\epsilon = 0.005$

Number of lines, n	*Probability of malfunction*
1,000	2.7×10^{-2}
2,000	2.6×10^{-3}
3,000	2.5×10^{-4}
5,000	4×10^{-6}
10,000	1.6×10^{-10}
20,000	2.8×10^{-19}
25,000	1.2×10^{-23}

the original system by the Scheffer-stroke system of $3n$ elements, each with its executive component and restoring component. The error in the over-all system can be computed from the error of the individual

Scheffer-stroke systems as shown in the previous discussions. Practically, this is a very tedious calculation. However, for the purpose of estimating the order of magnitude of n required for specified reliability, we can consider the entire system to be equivalent to a single Scheffer-stroke element, and use the result of this over-all reaction directly. This will be done in the following section.

18.6 Examples. To obtain an idea of the magnitude of required bundle size, let us consider a computing machine of 2,500 vacuum tubes. Each of the tubes is supposed to be actuated on the average of once every 5 microseconds. We specify that the machine should on the average run 8 hours before making a single mistake. During this period, the number of actuations of a single tube is

$$\tfrac{1}{5} \times 8 \times 3{,}600 \times 10^6 = 5.76 \times 10^9$$

Consider each tube as a Scheffer-stroke element. Then the specified probability of malfunction, considering each tube as an independent unit, is $1/(5.76 \times 10^9)$. However, there are 2,500 interconnected tubes in the system. A mistake in any one among the 2,500 tubes will mean a mistake of the machine. Therefore, considering each tube as one unit within the system, the specified probability of malfunction should be only 1/2,500 of the above value, or $1/(2{,}500 \times 5.76 \times 10)^9 = 7 \times 10^{-14}$. We see then that the final probability of malfunction is the same as that obtained by considering the whole system of 2,500 tubes as a single Scheffer-stroke element. This possibility greatly simplifies the calculation of required number of lines n in the multiplexed system.

If we assume that the fiduciary level δ and the probability ϵ of failure of the tubes are the same as specified in Table 18.1, then the specified probability of malfunction obtained above will require $n = 14{,}000$, according to the table. Therefore, in order to make the machine as reliable as specified, it will be necessary to multiplex the system 14,000 times. This would mean the replacement of every single tube in the machine by a system of $3n = 3 \times 14{,}000 = 42{,}000$ tubes. The original machine of 2,500 tubes now becomes a giant of 105,000,000 tubes. This is clearly not practicable.

Now take a second example, a plausible quantitative picture of the organization of the human nervous system. The number of neurons involved is usually given as 10^{10}. But considering the presence of synaptic end bulbs and other possible autonomous subunits, this number is certainly too low. It ought to be a few hundred times larger. Let us take the number of basic elements to be 10^{13}. The neurons can be actuated up to a maximum of about 200 times a second. But the average rate of actuations must be a good deal less; say 10 actuations per second. We shall further assume that a mistake in our nervous system is serious

and should not happen in the time interval of a human life span. Take the error-free interval to be 10,000 years. During this interval, the total number of actuations in the system of 10^{13} elements is

$$10^{13} \times 10{,}000 \times 31{,}536{,}000 \times 10 = 3.2 \times 10^{25}$$

Thus the probability of malfunction should be

$$1/(3.2 \times 10^{25}) = 3.2 \times 10^{-26}$$

Again, assume that the basic nervous elements have the properties specified in Table 18.1; then an extrapolation from that table gives $n = 28{,}000$.

However, our calculation needs a correction: if the human nervous system is indeed multiplexed 28,000 times, the number of basic elements in the nonmultiplexed system is not 10^{13} as assumed above; the number should be reduced by a factor of $1/(3 \times 28{,}000)$. Then the probability of malfunction should be increased by the factor $3 \times 28{,}000$. The corrected probability of malfunction is now 2.7×10^{-21}. Table 18.1 then gives $n = 22{,}000$. Further iteration will not change this value appreciably.

These examples show that, while our method of controlling the error by multiplexing is quite conceivably applicable to the microcomponents of nervous systems, it is nevertheless impractical for engineering control systems with the present technologies. One obvious direction of future development is the reduction in bulk and power requirement of the elements. The transistor is a great improvement over the vacuum tube from this point of view. Hence the method of multiplexing may yet become practical in the future. Another direction of investigation would be a deeper analysis of the process of controlling error. The organization of the executive component and the restoring component of our basic system for the Scheffer stroke is, after all, only one possible organization. We are fortunate in that such a crude attempt already is successful in demonstrating the possibility of increasing the reliability. There are probably other plans of organization of the duplicated components which will produce the same degree of reliability with a smaller number of components. In other words, only a beginning has been made in the technique of error control in automatic systems. For control engineers, there is as yet no applicable solution to the problem.

INDEX


116 638